ARI Research Note 88-56

Measuring the Vague Meanings of Probability Terms

Thomas S. Wallsten, Rami Zwick, and Barbara Forsyth
University of North Carolina at Chapel Hill

David V. Budescu and Amnon Rappaport
University of Florida

for

Contracting Officer's Representative
Michael Drillings

ARI Scientific Coordination Office, London
Milton S. Katz, Chief

Basic Research Laboratory
Michael Kaplan, Director

U. S. Army
Research Institute for the Behavioral and Social Sciences

July 1988

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

L. NEALE COSBY
Colonel, IN
Commander

Research accomplished under contract
for the Department of the Army

University of North Carolina

Technical review by

Dan Ragland

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS - -
2a. SECURITY CLASSIFICATION AUTHORITY - -	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE - -	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) - -	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Note 88-56
6a NAME OF PERFORMING ORGANIZATION University of North Carolina 6b. OFFICE SYMBOL (If applicable) - -	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute
6c. ADDRESS (City, State, and ZIP Code) Davies Hall, 013A Chapel Hill, CA 27514	7b. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600
8a. NAME OF FUNDING / SPONSORING ORGANIZATION - - 8b. OFFICE SYMBOL (If applicable) - -	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-83-K-0347
8c. ADDRESS (City, State, and ZIP Code) - -	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 6.11.02.B; PROJECT NO. 2Q1611 02B74F; TASK NO. n/a; WORK UNIT ACCESSION NO. n/a

11. TITLE (Include Security Classification)
MEASURING THE VAGUE MEANINGS OF PROBABILITY TERMS

12. PERSONAL AUTHOR(S) Thomas S. Wallsten, Rami Zwick, and Barbara Forsyth (University of North Carolina), David V. Budescu and Amnon Rappaport (University of Florida)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim Report	FROM Feb.85 TO Feb.86	June 1988	90

16. SUPPLEMENTARY NOTATION
Michael Drillings, contracting officer's representative

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Cognitive Psychology, Vague Meanings
Probability Theory, Judgement
Cognitive Science

19. ABSTRACT (Continue on reverse if necessary and identify by block number)
In two experiments, a modified pair-comparison procedure was employed in two experiments to establish and assess membership functions for numerous probability terms. In both cases, subjects judged: a) to what degree one probability term better described that probability than another, and b) to what degree one term rather than another better described a probability. Task a) data from subjects was analyzed in terms of the axioms of an algebraic-difference structure, and membership function values were obtained for each term according to various ratio and difference scaling models. The axioms were well satisfied, and goodness-of-fit measures for the scaling procedures were quite high. Furthermore, the derived membership functions had interpretable shapes and satisfactorily predicted for each subject the judgements independently obtained in b). These results support the claims that the scaled values indeed represented the vague meanings of the terms to the subjects in the present context.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL Thomas S. Wallstein	22b. TELEPHONE (Include Area Code) - -
	22c. OFFICE SYMBOL - -

DD Form 1473, JUN 86 Previous editions are obsolete.

MEASURING THE VAGUE MEANINGS OF PROBABILITY TERMS

Thomas S. Wallsten

University of North Carolina at Chapel Hill

David V. Budescu Amnon Rapoport

University of Haifa

Rami Zwick Barbara Forsyth

University of North Carolina at Chapel Hill

Abstract

Many authors have suggested that the vague meanings of probability terms such as doubtful, probable, or likely, can be expressed as membership functions over the [0,1] probability interval. A function takes value zero for probabilities not at all in the vague concept represented by the term, one for probabilities definitely in the concept, and intermediate values otherwise. A modified pair-comparison procedure was employed in two experiments to empirically establish and assess membership functions for numerous probability terms. In both cases, subjects (graduate students in the social sciences and business) judged (a) to what degree one probability rather than another was better described by a specific probability term, and (b) to what degree one term rather than another better described a specific probability. Probabilities were displayed as relative areas on spinners. Task (a) data from individual subjects were analyzed in terms of the axioms of an algebraic-difference structure, and membership function values were obtained for each term according to various ratio and difference scaling models. The axioms were well satisfied and goodness of fit measures for the scaling procedures were quite high. Furthermore, the derived membership functions had interpretable shapes and satisfactorily predicted for each subject the judgments independently obtained in (b). These results support the claim that the scaled values indeed represented the vague meanings of the terms to the subjects in the present context. Subjects' membership functions were stable over time, but except for the term tossup showed large individual differences. The data are discussed in terms of both their methodological and substantive implications.

The procedures developed here may be especially useful in subsequent research on the factors that affect the meanings of probability terms and on how vague uncertainties are processed. In addition, they can easily be applied in other semantic domains as well.

MEASURING THE VAGUE MEANINGS OF PROBABILITY TERMS

Most people, including expert forecasters, generally prefer communicating their uncertain opinions with nonnumerical terms such as doubtful, probable, slight chance, very likely, and so forth, rather than with numerical probabilities. On anecdotal grounds, the imprecision of nonnumerical terms is preferred to the precision of probability numbers for at least two reasons: First, opinions are generally not precise and therefore, the claim goes, it would be misleading to represent them precisely. For example, commenting that numbers denote authority and a precise understanding of relationships, a committee of the U.S. National Research Council wrote that there is an

> ...important responsibility not to use numbers, which convey the impression of precision, when the understanding of relationships is indeed less secure. Thus, while quantitative risk assessment facilitates comparison, such comparison may be illusory or misleading if the use of precise numbers is unjustified (National Research Council Governing Board Committee on the Assessment of Risk, 1981, p. 15).

The second reason frequently suggested for communicating with nonnumerical terms rather than with probability numbers is that most people feel they better understand words than numbers. Zimmer (1983) pointed out that it was not until the 17th century that probability concepts were formally developed, yet expressions for different degrees of uncertainty existed in many languages long

before then. He (Zimmer, 1984) suggested that people generally handle uncertainty by means of verbal expressions and their associated rules of conversation, rather than by means of numbers.

The dual claims that vague opinions are well communicated with probability expressions and that people more naturally think about uncertainty in a verbal than in a numerical manner, can be investigated only after methods have been developed for validly measuring the vagueness associated with probability terms. Recognizing that the meanings of words are subject to individual differences and numerous context factors, the present research is primarily methodological and exploratory, aimed at developing suitable measurement techniques and at making preliminary statements about probability terms. If procedures for validly measuring vagueness can be established, they can be employed to investigate the many substantive issues.

In most of the empirical work to date on the meaning of probability words, subjects have been asked to give numerical equivalents to various probability phrases. The overwhelming result has been that there is great intersubject variability in the numerical values assigned to probability terms and great overlap among terms (Bass, Cascio & O'Connor, 1974; Beyth-Marom, 1982; Budescu & Wallsten, in press; Foley, 1959; Johnson, 1973; Lichtenstein & Newman, 1967; Simpson, 1944, 1963). Within-subject variability in the assignment of numbers to probabilistic terms is not minor, but is considerably less than between-subject variability (Beyth-Marom, 1982; Budescu & Wallsten, in press; Johnson, 1973). However, neither the within- nor the between-subject variability alone can be taken as evidence that probability terms have vague

meanings. First of all, as pointed out by Budescu and Wallsten (in press), there is no way to determine whether the variability is due to differences between subjects, or within subjects over time, in the use of numbers rather than in the use of words. Secondly, and more to the present point, as Rubin (1979) noted in a related context, these data can as well be interpreted as showing that the meanings of probability terms are not constant over people or times as showing that the expressions have generally vague meanings. Thus, an alternative approach is necessary.

Membership Functions

Numerous authors (e.g., Watson, Weiss, & Donnell, 1979; Zadeh, 1975; Zimmer, 1973) have suggested that the meaning of a probability term can be represented by a function on the [0,1] probability interval, as illustrated in Figure 1. The function takes its minimum value, generally zero, for probabilities that are not at all in the concept represented by the phrase, its maximum value, generally one, for probabilities definitely in the concept, and intermediate values for probabilities with intermediate degrees of memberships in the concept represented by the term. There are no constraints on the shapes such functions can have, nor must they be expressable by equations of any particular sort. Within fuzzy set theory, such a function is called a membership function, but it is not necessary to tie this idea strictly to fuzzy set theory.

Insert Figure 1 about here

Of course, the question of defining and measuring the vague

meaning of a term arises in a vast array of semantic domains, and the concept of a membership function has been applied quite broadly within fuzzy set theory (e.g., Norwich & Turksen, 1984; Zadeh, 1975; Zysno, 1981). As a general definition, a membership function is a rule that assigns to each element in the universe of discourse a number in the closed [0,1] interval indicating the degree to which that element is a member of a particular set or category. If the category is well defined (e.g., male humans beyond their 60th birthday), then all membership functions are either 0 or 1. If the category is not well defined (e.g., middle aged men), then the membership functions can take on any value in the [0,1] interval.

Measurement of Vagueness

A considerable literature exists on the topic of vagueness (e.g., Ballmer & Pinkal, 1983; Gaines & Kohout, 1977; Goguen, 1969; Hemple, 1939; Hersh & Caramazza, 1976; Labov, 1973; Oden, 1981; Skala, Termini, & Trillas, 1984; Zadeh, 1965). However, although much has been written about the measurement of vagueness or fuzziness, empirical work has been relatively sparse. One method relies on choice probabilities. For example, a stimulus, such as a square, is presented along with a word such a small (Hersh & Caramazza, 1976; Hersh et al., 1979). The subject answers yes or no according to whether the word describes the stimulus. The fraction of yes responses over subjects or within subjects over trials is then taken as the degree of membership for that stimulus in the vague concept represented by the word. Rubin (1979) has criticized this procedure because (a) it confounds measures of fuzziness with response variability due to experimental procedures, and (b) it can just as well be interpreted as showing that words have different

meanings to different people or at different times as that words have vague or fuzzy meanings.

A second method of obtaining membership functions is direct scaling, in which subjects rate stimuli on a scale from "definitely in the concept" to "definitely not in the concept." For example, Oden (1977a) had subjects rate propositions on a scale from absolutely true to absolutely false and Zysno (1981) had subjects rate grade of membership on a scale from 0% to 100% of a man X years of age in concepts such as old man, very young man, etc., for various values of X (see also MacVicar-Whelan, 1978). In other studies (e.g., Kuz'Min 1981), subjects picked stimuli with specified grades of membership. The direct scaling methods overcome some of the problems with the choice probabilities, in that the construct of vagueness is directly assessed in individual responses. However, as with all magnitude estimation procedures, the responses cannot be evaluated unless they are embedded within a theory. Oden used functional measurement techniques to assess his measures; many other authors simply display the estimates after they are obtained (e.g., Norwich & Turksen, 1984) or fit them with explicit functions which are evaluated by means of goodness of fit measures (e.g., Zysno, 1981).

We employ a different approach which utilizes a modified pair comparison method for measuring the vagueness of probability terms. Empirically, the procedure is similar to one utilized by Oden (1977b), but the data are analyzed much differently. The data can be first checked at an ordinal level to determine if they satisfy certain axioms necessary for scaling vagueness according to an

algebraic difference (or ratio) model (Krantz, Luce, Suppes, & Tversky, 1971). If the axioms are reasonably well satisfied, then specific difference or ratio scaling procedures (Saaty, 1977, 1980; Torgerson, 1958) can be applied to the data for the purpose of deriving the vagueness measure, or membership function, for each expression. Furthermore, goodness of fit measures can be calculated to evaluate the quality of the metric scaling.

Insert Figure 2 about here

A Pair-comparison Method

To make the discussion concrete, consider a sample experimental trial as shown in Figure 2. Two spinners are drawn on a computer monitor. Subjects are told to imagine a pointer over each spinner that can be spun so that it randomly lands over either the white or the dark sector. Thus, each spinner displays a different probability of the pointer landing on white. There is a probability term printed above the spinners and a line with an arrow on it below them. The subject must move the arrow on the line to indicate for which spinner the probability of landing on white is better described by the probability term and how much better it is described. Moving the arrow to the far left indicates that the left spinner is absolutely better described, leaving the arrow in the middle indicates that the two spinners are equally well described, and so forth. The probabilities on the two spinners are changed from trial to trial according to a left side by right side, $\underline{P} \times \underline{P}$, factorial design in which $\underline{P} = \{\underline{p}_1,\ldots,\underline{p}_n\}$, where for $\underline{i} = 1,\ldots,\underline{n}$, the $\underline{p}_i$ denote specific probabilities of the spinners

landing on white.

Consider the bounded response line shown in Figure 2 to extend from 1 on the left to 0 on the right and let $\underline{R}_W(\underline{ij})$ be the response when probability $\underline{p}_i$ is on the left, $\underline{p}_j$ is on the right, and expression $\underline{W}$ is displayed above them. The responses $\underline{R}_W(\underline{ij})$ induce an ordering on the factorial design according to the degree that the left hand probability is better described by the term than is the right hand probability. If, as will be described, this ordering satisfies the axioms of an algebraic difference structure (Krantz, et al., 1971), then a suitable transformation of the cell entries can be used in a difference or a ratio scaling model to establish a membership function for the term $\underline{W}$, such as is shown in Figure 1.

A bit of notation will aid in making these concepts clear. Let $(\underline{p}_i,\underline{p}_j)$ refer to a cell in the $\underline{P}$ x $\underline{P}$ factorial design, or in other words, be an element in the Cartesian product of $\underline{P}$ x $\underline{P}$. The elements of the Cartesian product, or the cells of the factorial design, are rank ordered according to how much better phrase $\underline{W}$ describes the lefthand probability than the righthand probability. The rank ordering between any pair of cells is denoted by $\succsim_W$ where the subscript indicates that the ordering is for the particular phrase (doubtful in Figure 2). As indicated above, the ordering is induced by the placement of the arrow on the response line, so that the further to the left an arrow is for a pair of probabilities, the higher in the rank ordering is that pair. Stated formally,

$$p_i p_j \succsim_W p_k p_l \quad \text{iff } R_W(ij) \geq R_W(kl)$$

Let $(\underline{P}x\underline{P}, \succsim_W)$ refer to an ordered matrix of the sort just described. Krantz et al.(1971) prove that if $(\underline{P}x\underline{P}, \succsim_W)$ satisfy

five axioms, then there exists a mapping μ_W from $\underline{P}$ into the real numbers such that

$$p_i p_j \gtrsim_W p_k p_l \text{ iff } \mu_W(p_i) - \mu_W(p_j) \geq \mu_W(p_k) - \mu_W(p_l),$$

or, equivalently, such that

$$p_i p_j \gtrsim_W p_k p_l \text{ iff } \mu_W(p_i) / \mu_W(p_j) \geq \mu_W(p_k) / \mu_W(p_l).$$

In other words, scale values can be assigned to these probabilities such that the rank order of differences (or of ratios) in the assigned values matches the rank order of differences (or of ratios) in the degrees to which the lefthand and righthand probabilities are described by the phrase. The scale values are unique up to a linear (for the difference representation) or a power (for the ratio representation) transformation. These scale values, normalized to be nonnegative with an arbitrary maximum of 1, and plotted as a function of the probabilities (as illustrated in Figure 1) can be taken as the membership function representing the degree to which each probability belongs to the vague concept defined by the expression.

It should be noted that at an axiomatic level, the difference and ratio representations cannot be distinguished unless different orderings appear under difference- and ratio-inducing conditions (see Birnbaum, 1980, and Miyamoto, 1983). This is because any set of differences can be mapped into a set of ratios by taking logs, and conversely, any set of ratios can be mapped into a set of differences by exponentiating.

Tests of the Axioms

The five axioms specified by Krantz et al. (1971) include two that are of purely mathematical interest and three that can be

subjected to empirical test. One of these, the weak order axiom, states that the elements of P x P all can be compared to each other and that the ordering is transitive. Our method of using the arrow location to rank order the matrix forces this axiom to be satisfied and therefore it is not of empirical interest here. However, the remaining two axioms, sign reversal and weak monotonicity, can be evaluated.

The weak monotonicity axiom is illustrated in Figure 3. It states that for all $\underline{p}_i$, $\underline{p}_j$, $\underline{p}_k$, $\underline{p}_{i'}$, $\underline{p}_{j'}$, and $\underline{p}_{k'} \in \underline{P}$, if $\underline{p}_i\underline{p}_j \succsim_W \underline{p}_{i'}\underline{p}_{j'}$ and $\underline{p}_j\underline{p}_k \succsim_W \underline{p}_{j'}\underline{p}_{k'}$, then $\underline{p}_i\underline{p}_k \succsim_W \underline{p}_{i'}\underline{p}_{k'}$. Single arrows in Figure 3 indicate the antecedent conditions and the double arrow indicates the consequent.

Insert Figure 3 about here

The monotonicity axiom can be evaluated separately within the P x P matrix associated with each term. This is done by selecting suitable subsets of six cells within the matrix and then for all those subsets for which the antecedent conditions are met, checking to determine whether the consequent condition is also met. The number of subsets available for test depends on the size of the matrix and can be substantial. Of course, there is considerable overlap among the subsets, and therefore the tests are not independent. A convenient summary statistic for each matrix is the percentage of possible tests that are satisfied.

The sign reversal axiom states that for all $\underline{p}_i$, $\underline{p}_j$, $\underline{p}_k$, and $\underline{p}_l \in \underline{P}$, if $\underline{p}_i\underline{p}_j \succsim_W \underline{p}_k\underline{p}_l$, then $\underline{p}_l\underline{p}_k \precsim_W \underline{p}_j\underline{p}_i$. The axiom is checked easily on all suitable quadruples of cells.

Norwich and Turksen (1982, 1984) were apparently the first to recognize the close relationship between the axiomatic formulation of the algebraic difference structure and the validation of membership functions. They provide an elegant mathematical development of the measurement system just outlined. Strangely, for their accompanying experiment (Norwich & Turksen, 1984), they merely state that the axioms are satisfied without presenting supporting data and then use a simple magnitude estimation procedure to establish the membership functions. However, in the absence of additional strong assumptions there is no necessary relation between scale values obtained by magnitude estimation and those obtained by a pair comparison procedure.

It is important to note that a pure pair-comparison procedure will yield ordinal data sufficient for checking the axioms and also for nonmetric scaling, but will not provide data from which membership functions can be derived by means of metric scaling procedures. The present modified or any other graded pair comparison method (Sjöberg, 1980) does yield data that can be analyzed in terms of both axiomatic and metric models.

Scaling Models

One approach to applying the metric scaling models proceeds as follows. Consider the difference model first. Assume that for a given expression W and probability pair $p_i p_j$, the subject places the arrow on the response line such that the difference in the distances of the arrow from the two ends is inversely proportional to the difference in the degrees to which W describes p_i and p_j. Thus, the response R can be converted to a difference score D for purposes

of scaling:

$$(1) \qquad D_W(ij) = 2R_W(ij) - 1.$$

The proportionality assumption plus an assumed error component yield

$$(2) \qquad D_W(ij) = \alpha_W[\mu_W(p_i) - \mu_W(p_j)] + \varepsilon_{Wij},$$

with $\alpha_W > 0$. Considering the full matrix of difference scores for phrase $\underline{W}$, a least squares estimate of $\mu_W(\underline{p}_i)$ is obtained by taking row means. In other words, from Equation (2),

$$(3) \qquad \hat{\mu}_W(p_i) = \sum_j D_W(ij) / n$$

where $\underline{n}$ is the size of the matrix and $\alpha_W = 1$. The scale values, of course, are unique up to a linear transformation, $\mu_W'(\underline{p}_i) = \alpha_W\mu_W(\underline{p}_i) + \beta_W$, and can easily be rescaled to be positive with a maximum at 1. Note that the scaling is done independently for the $\underline{P}$ x $\underline{P}$ matrix associated with each $\underline{W}$. Thus membership values across phrases are not comparable without additional assumptions.

For the ratio scaling models, it is assumed that the arrow is placed on the response line such that the ratio of the distances of the arrow from the two ends is inversely proportional to the ratio of the degrees to which $\underline{W}$ describes $\underline{p}_i$ and $\underline{p}_j$. Thus, the response $\underline{R}$ is converted to a ratio score $\underline{S}$:

$$(4) \qquad S_W(ij) = R_W(ij) / [1 - R_W(ij)],$$

with $\underline{R}_W(\underline{ij}) \neq 0,1$. Now the proportionality assumption plus assumed error yields

$$(5) \qquad S_W(ij) = \alpha_W \varepsilon_{Wij} \mu_W(p_i) \; / \; \mu_W(p_j) \; ,$$

with $\alpha_W > 0$. The geometric means,

$$(6) \qquad \hat{\mu}_W(p_i) = [\pi_j \; S_W(ij)]^{1/n} \; ,$$

with $\alpha_W = 1$, are least squares estimates of the logarithms of the scale values (Torgerson, 1958). The resulting scale values are unique up to a power transformation, $S_W'(\underline{p}_i) = \alpha_W S_W(\underline{p}_i)^{\beta W}$, with $\alpha_W, \beta_W > 0$.

An alternative ratio scaling procedure, anticipated by Gulliksen (1958) and developed by Saaty (1977, 1980), requires a reciprocal matrix, i.e., one in which $\underline{S}_W(\underline{ij}) = 1/\underline{S}_W(\underline{ji})$. Scale values can be obtained from the matrix by taking row geometric means (GM) or by an eigenvalue-eigenvector decomposition, obtaining either a normalized right eigenvector (RE), a normalized left eigenvector (LE), or the mean of the two eigenvectors (ME). If a reciprocal matrix is consistent (i.e., for any three entries, $\underline{S}(\underline{ij})$, $\underline{S}(\underline{jk})$, and $\underline{S}(\underline{ik})$, $\underline{S}(\underline{ik}) = \underline{S}(\underline{ij})\underline{S}(\underline{jk})$), then GM, RE, LE, and ME all yield the same scales. Otherwise they do not, and there is currently some controversy concerning the merits of each solution. Properties of the various solutions have been investigated mathematically (e.g., De Jong, 1984; Jensen, 1984; Saaty and Vargas, 1984) and with Monte Carlo procedures (e.g., Budescu, Zwick, and Rapoport, 1985; Johnson,

Beine, and Wang, 1979; Williams and Crawford, 1980). However, the four methods have not been compared on real data, so it appears premature to reject one in favor of the others.

Saaty (1977, 1980) has proposed a goodness of fit index that compares the maximum eigenvalue to the size of the matrix. However, a more general goodness of fit measure that allows the four ratio and the difference scaling models all to be compared is the linear correlation between the observed and the predicted responses. Therefore, we will employ this measure to evaluate the metric scaling models.

Cross Validation

It is necessary for the validity of any of these models that the axioms be satisfied within limits of error and that the model's goodness of fit measure be high. However, such tests are not sufficient for supporting the more interesting claim that the vague meanings of the expressions are validly represented by the derived scale values. For example, this claim would appear unjustified if the scale values plotted as a function of the probabilities (cf. Figure 1) yielded uninterpretable curves, e.g., multipeaked. Furthermore, for this claim to be justified, it is necessary that the derived values correctly predict an independent set of judgments based on the presumed vagueness of the terms.

In the present experiments, subjects were also run on trials that were the converse of that shown in Figure 2; namely, there was one spinner at the top of the screen with two terms below it, one on the left and one on the right. The subject moved the arrow on the response line to indicate how much better one term rather than the

other described the displayed probability of landing on white. Scale values derived from the previous judgments should predict certain properties of these responses. The predictions are derived here only in terms of scales obtained from Equation (6), because ultimately those were the values with which they were tested.

Consider an experimental trial with probability $\underline{p}$ and terms $\underline{W}_i$ and $\underline{W}_j$, for which the subject sets the arrow at location $\underline{R}_p(\underline{ij})$. (Note the shift in notation to correspond with the change in the structure of a trial. We are now assuming a fixed $\underline{p}$ and a set of terms $\underline{T} = \{\underline{W}_1, \ldots, \underline{W}_m\}$.) The response value $\underline{R}_p(ij)$ is transformed to $\underline{S}_p(\underline{ij})$ by Equation (4) (with indices suitably changed). If the previously derived scale values $\mu_{W_i}(\underline{P})$ and $\mu_{W_j}(\underline{P})$, represent the degree to which $\underline{p}$ is a member of $\underline{W}_i$ and $\underline{W}_j$, respectively, then it should be the case that

$$(7) \qquad S_p(W_iW_j) = \delta \, \mu_{W_i}(P)^{\beta_i} / \mu_{W_j}(P)^{\beta_j}$$

where $\delta, \beta_i, \beta_j > 0$. For clarity, the scaling parameters are not fully subscripted. But they have been included in Equation (7), because it is important to note what assumptions are being made about them.

Consider first a fixed pair of phrases $\underline{W}_i$ and $\underline{W}_j$ and various $\underline{p}$, all of which have non-zero membership functions in $\underline{W}_i$ and $\underline{W}_j$. If it is assumed that $\beta_i = \beta_j = 1$, then from Equation (7) the $\underline{S}_p(\underline{W}_i\underline{W}_j)$ should be a linear function of the ratios of the derived membership functions. This prediction was tested in Experiment 1.

Now consider a fixed probability p with various phrases W_1, $\underline{W}_2,\ldots,\underline{W}_m$. In this case, $\underline{S}_p(\underline{ij})$ is a linear function of the ratio of the derived membership functions only if it is assumed that $\delta = \beta_i = \beta_j =$ for all W_i and W_j. This is tested in Experiment 2.

A very strong prediction emerges if for a given $\underline{p}$ there is a left side x right side $\underline{T}$ x $\underline{T}$ factorial design, in which $\underline{T}$ is the vector of probability terms. The data matrix for each $\underline{p}$ can be scaled in a manner analogous to that described with Equations (4) - (6). The resulting scale values, $\alpha_p(\underline{W})$, are unique up to a power transformation. Omitting subscripts, on the reasonable assumption that $\mu_p(\underline{W})$ and $\mu_W(\underline{P})$ both represent the same vagueness construct, it is easy to show that the two sets of derived values should be related by a power function,

$$\mu_p(\underline{W}) = \alpha\mu_W(\underline{P})^{\beta}, \tag{8}$$

with $\alpha, \beta > 0$. This, too, is tested in Experiment 2.

While the various empirical evaluations could be carried on in many domains, the present experiments do so for the vague concepts defined by probability expressions. Specifically, the purposes of the present experiments are (a) to evaluate the measurement models by testing their ordinal and goodness of fit predictions, (b) to evaluate the claim that the derived values represent the vague meanings of the phrases boty testing their ordinal and goodness of fit predictions, (b) to evaluate the claim that the derived values represent the vague meanings of the phrases both by considering the reasonableness of the resulting

scales and by predicting an independent set of judgments, and (c) to make some preliminary statements about meanings of nonnumerical probability expressions.

Experiment 1

Two groups of subjects were employed, each responding to a different set of probability terms, as shown in Table 1. A simple context phenomenon that could be investigated in this study was whether the derived membership function for a term depended on the set of terms under consideration. Therefore, Group 1 had terms weighted toward the high end of the probability continuum, and Group 2 had terms weighted toward the low end with, however, six words in common.

Insert Table 1 about here

Subjects were run in one session for practice followed by two for data. As shown in Table 1, different terms were used for the two data sessions in order to increase the number of terms employed. However, the term possible was utilized on both days to get some notion of the stability over time of the membership function.

We considered the experimental task to be a difficult one, and therefore made a number of decisions intended to maximize the quality of the data. First, we elected to use social science and business graduate students rather than undergraduates as subjects. We assumed that they would represent a population of people who think seriously about communicating degrees of uncertainty, and who generally do so with nonnumerical phrases.

Second, the probabilities used with each term were determined

uniquely for each subjet. Furthermore, each probability pair was presented only once with a given term in a session. Thus, in terms of the data entries illustrated in Table 1, if probability p_i was presented on the left and p_j on the right, the arrow location ($R_W(ij)$, expressed as a number from 1 to 0) was entered in cell ij and its complement ($1-R_W(ij)$) was entered in cell ji. While this procedure has statistical drawbacks, it greatly reduced the number of trials and the motivation for subjects to hurry through the session. Of particular interest to this study, the procedure forced the sign reversal axiom to be correct (leaving only weak monotonicity to be evaluated), and also yielded the reciprocal matrix required by Saaty's ratio scaling technique.

Method

Subjects. Subjects were recruited by placing notices in graduate student mailboxes in the business school and the departments of anthropology, economics, history, psychology, and sociology. The general nature of the study was described and subjects were promised $25 for three sessions of approximately an hour and a half each. Ten native speakers of English were randomly assigned to each of Groups 1 and 2. As explained in conjunction with Table 1, the groups differed only in terms of the words they judged.

General procedure. Subjects were run for a practice and then two data sessions, with the sessions scheduled generally two days apart. The experiment was controlled by an IBM PC with stimuli presented on a color monitor and responses made on the keyboard. During the practice session, all subjects judged the phrases chance,

very likely, and slight chance. During Sessions 2 and 3, subjects judged terms as indicated in Table 1. An index card was continuously in view listing all the expressions that the subject would encounter during the course of the experiment.

Each session consisted of three parts. The purpose of Part 1 was to determine the maximum, p^*, and the minimum, p_*, probability for which the subject would judge a given term to be appropriate. The results of this part were then used to determine the unique probabilities to be employed in Parts 2 and 3 for each subject.

The second part of the session involved the presentation of probability terms with pairs of spinners, as already discussed. Part 3 reversed the procedure, as also already discussed. Each part will now be described in more detail.

Part 1. The instructions for this segment read in part: In a specific context that we will describe shortly, we are interested in the range of uncertainties for which you think it appropriate to use each of various words or phrases that will be displayed on the screen ...

The context that we will provide is that of spinning a pointer on a spinner that is radially divided into a red sector and a white sector. The relative areas of each sector are clear to you and you must convey that information to a friend. You want to tell him how likely it is that the pointer will land on white if it is fairly spun and randomly stops at some position. However, you are not allowed to tell the person the actual probability of landing on white. Rather, you are forced to use a nonnumerical descriptive phrase ... We want to know the range of probabilities in

> this specific spinner context for which you would consider (each term) to be appropriate ...

The terms scheduled for a given session were presented in random order. On each trial a phrase was written at the top of the screen and a spinner divided vertically into equal areas of red and white was drawn below it. The subject then increased the proportion of white by pressing the I key and decreased it by pressing the D key. The relative area of white was first adjusted to indicate the lowest probability for which the subject would conceivably use the displayed term. This value was then registered by pushing the L key. The instructions for this task read in part:

> ... Adjust the spinner to some low probability of landing on white and ask yourself, "Would I conceivably apply (the displayed term) to that probability?" If the answer is yes or possibly, then set the spinner to a lower probability and ask yourself the same question again. If the answer is definitely not, then increase the probability a little and repeat the question. Continue in this fashion until you achieve the very lowest probability to which you might apply (the term). That is your lower limit.

After the lower limit was indicated, the subject then adjusted the spinner to display the highest probability for which he or she might use the term, which was registered by pressing the U key. The upper limit could not be set below the lower limit.

Instructions for this part ended with three reminders: 1. To consider the use of the expression only in terms of describing the chances of the pointer landing on white for the particular spinner

displayed on the screen, not with how it might be used in other contexts; 2. Not to decide whether the particular term is the best of all possible terms for a given probability, but only whether it could conceivably apply to the displayed relative area; and 3. To select the lowest and highest probabilities carefully, because they were to be used to determine the range of probabilities employed with each expression in the subsequent parts of the experiment.

Immediately following Part 1, the interval from $\underline{p}_*$ to $\underline{p}^*$ for each term was divided online into $\underline{n}$ equally spaced probability values for use in Part 2. For each term, $\underline{n}$ was set at the largest integer between 0 and 8, inclusive, such that the spacing of adjacent probability values was not less than 0.02.

Part 2. Depending on the Part 1 results, the number of probabilities, $\underline{n}$, associated with each term ranged from 0 to 8. Terms were presented in this part only if $\underline{n} \geq 2$. Probabilities were displayed as the relative areas of white on a spinner. Each phrase was presented once with each of the $\underline{n}(\underline{n}-1)/2$ pairs of spinners. Phrases and spinner pairs were presented in a random, not a blocked order.

A single trial appeared as shown in Figure 2. As already indicated, the subject moved the arrow on the line to indicate for which spinner the probability of landing on white was better described by the expression and how much better it was described.

The instructions said in part:

> ... If you had to assign the phrase at the top of the screen to one of the two spinners, to describe the probability of landing on white, to which spinner is it more appropriately assigned and how much more appropriate is the assignment of

> the phrase to that spinner than to the other one? ... If you believe the two probabilities are equally well described by the phrase, leave the arrow in the middle. If the probability on one spinner is better described by (the term) than is the other, move the arrow closer to that spinner. The greater the relative appropriateness of the phrase for one probability than for the other, the closer the arrow should be moved to the corresponding spinner. In other words, place the arrow so that its relative distance between the two spinners represents its relative appropriateness for the two probabilities.

The < and > keys on the keyboard were used to move the arrow on the screen, the R key was used to register the response when the arrow was suitably placed. The arrow could be positioned at any of 17 equally spaced locations on the line, consistent with response procedures normally used for Saaty's (1977, 1980) ratio scaling techniques.

Part 3. This was the converse of Part 2. A pair of terms was presented only if the Part 1 estimates for the two terms overlapped. During Session 2, pairs were selected only from terms that were employed in Parts 1 and 2 of that session. Pairs were selected the same way in Session 3, but in addition, pairs were formed with one member from Session 2 and one from Session 3 if their Part 2 estimates overlapped sufficiently. The number of probabilities presented with a pair ranged from 1 to 8, with adjacent probabilities differing by at least 0.02. Due to a programming error, the Session 2 and 3 presentations of possble were treated

separately. Thus, in Session 3, possible may have been paired with other phrases up to 16 times each. Spinner and phrase pairs were presented in a random, not a blocked order.

On a trial, a spinner representing a particular probability was presented at the top of the screen; two terms were written below it, and a marked line segment with a centered arrow was below them. In the same manner as in Part 2, the subject moved the arrow on the line segment to indicate which of the two terms better described the probability of the spinner landing on white and how much better the description was.

The instructions read in part:

> If you had to select one of the two phrases to describe the displayed probability of landing on white, which of the two is better, and relatively how much better is it? ... The relative distance you place the arrow between the two phrases should represent relatively how much better one phrase is for the displayed probability than is the other.

Results

Virtually all analyses were done on individual, not group data. As will be documented subsequently, no apparent differences emerged between the two groups, so the group distinction will be generally disregarded. Data will be presented separately for the three parts of the experiment.

Part 1. Each subject set upper and lower limits for the range of probabilities that could be associated with each expression. A summary over subjects of these estimates is shown in Figure 4. For each term, the lower lefthand bar shows the interquartile range of the lower limit determinations. That is, the bar extends from the

25th percentile to the 75th percentile of the judgments over subjects. Similarly, the lower righthand bar indicates the 25th and 75th percentiles of the upper limit determinations. The medians of the lower and the upper limit determinations are connected by the top bar for each term. Note (a) the considerable variability over subjects, (b) that even the word tossup has a range of meanings from about 0.4 to 0.6 for most subjects, and (c) the enormous differences over subjects in the range of values suitable for the word possible.

Insert Figure 4 about here

Despite the considerable between-subject variability in the upper and lower limits for possible, individual subjects were reasonably stable over sessions. The correlation over subjects between the first and second determinations of the lower limit for possible was 0.94 ($p < 0.0001$). The correlation for the upper limit was 0.69 ($p < 0.001$).

Part 2. Each of 20 subjects set upper and lower limits for 9 expressions (counting possible separately for Sessions 2 and 3), for a total of 180 determinations. The width of each interval determined the number of probabilities to be associated with the corresponding term in this part. At most eight probabilities were selected to be equally spaced within the interval such that adjacent values differed by at least 0.02. Thus, intervals that were at least 0.16 wide yielded 8 probability values, while, for example, intervals that were at least 0.06 but less than 0.08 yielded 3 probability values. Table 2 shows the frequency distribution over interval

size, $\Delta p = p^* - p_*$, and over the corresponding derived number of probabilities, n, that were used in Part 2. It can be seen in the table, for example, that on 144 occasions, 8 probability values were associated with terms in Part 2, on 7 occasions 7 probabilities were associated with terms, and so forth. On eight of the 180 determinations, subjects set the upper and lower probability limits equal to each other, resulting in zero probability values to be used with the corresponding term, and therefore, in that term never appearing in Parts 2 or 3.

Insert Table 2 about here

The data for Part 2 were analyzed with respect to three questions. First, are ordinal properties of the judgments for a given term consistent with the axioms of an algebraic difference structure? Second, are metric properties of the judgments well described by one or more of the scaling models? Third, do the resulting membership functions have reasonable shapes?

Considering the ordinal data properties first, judgments were collected in this experiment in a manner such that both the weak ordering and the sign reversal axioms were forced to be satisfied. However, the weak monotonicity axiom could be tested.

Evaluation of the axiom required a matrix of size $n > 4$. Because only one of each reciprocal pair of cells in the $P \times P$ matrix for a phrase was responded to, the number of subsets of six cells for which the axiom could be tested equaled $\binom{n}{3}^2 - \binom{n}{3}$. A satisfaction index, defined as the percentage of subsets satisfying the consequent condition that also satisfied the antecedent

conditions, was determined for each phrase for which $n > 4$ for each subject.

The results of the weak monotonicity test are summarized in Table 3. Since the properties of this test and our summary statistic (percent of tested subsets satisfying the axiom) are not known, the test was applied to 400 random matrices of each size encountered in this experiment. The mean and standard deviation for percent satisfaction for the random data are shown in the top part of the table. There appears to be no effect of matrix size on the mean percent of subsets satisfying weak monotonicity, with an overall mean value of 54.5%. The standard deviation appears to decrease with matrix size.

Insert Table 3 about here

The bottom portion of the table summarizes the actual data as a function of matrix size. Each column in the table first shows the number of matrices analyzed. There was, of course, a distribution of satisfaction indices at each matrix size, and the subsequent entries in the columns of the table show the 25th, 50th, and 75th percentiles of those distributions. Taking a weighted average over matrix size, 75 percent of the matrices had satisfaction indices greater than 75.2%, 50 percent of the matrices had satisfaction indices greater than 82.3%, and 25 percent of the matrices had satisfaction indices that exceeded 89.4%. From another perspective, the last row of the table shows the percentage of matrices at each size that had satisfaction indices exceeding the mean value for

random data by at least three standard deviations. It seems reasonable to conclude that weak monotonicity is well satisfied.

We now turn to the metric scaling. For this purpose, the 17 equally spaced response locations were assigned values from left to right of 1, 0.9375, ..., 0.0625, 0. Then, since subjects responded to only one member of each pair of reciprocal cells in a matrix, the complementary response was entered in the other cell. That is, if $\underline{R}_W(\underline{ij})$ was the response to $(\underline{p}_i,\underline{p}_j)$ for phrase $\underline{W}$, $\underline{R}_W(\underline{ji}) = 1 - \underline{R}_W(\underline{ij})$ was entered in cell $(\underline{p}_j,\underline{p}_i)$.

Each matrix was scaled according to the difference model through application of Equations (1) and (3). In order to transform responses by Equation (4) for ratio scaling, responses, $\underline{R}_W(\underline{ij})$, of 0 and 1 were first set equal to 0.0156 and 0.9844, respectively (i.e., 1/4 of the distance between the most extreme and the immediately adjacent responses), to avoid division by 0. Then the geometric mean ratio scaling was accomplished via Equation (6). Ratio scaling solutions were also obtained by a right eigenvector-eigenvalue decomposition, a left eigenvector-eigenvalue decomposition, and by taking the means of the two eigenvectors.

All five scaling models were evaluated in terms of the linear correlation between observed and predicted responses. The mean correlations over all subjects and phrases were 0.75, 0.77, 0.75, 0.75, and 0.76 for the difference, geometric mean, right eigenvector, left eigenvector, and mean eigenvector models, respectively. Thus, all the models scaled the data about equally well, as though the geometric mean model is slightly superior on the average. Detailed results will be presented only for the geometric mean model; the others show similar patterns.

Recall from Table 2 that on eight occasions the upper and lower probability limits from Part 1 were set equal to each other so that the phrases did not appear in Part 2. Thus, 172 matrices were scaled, and for each a linear correlation was calculated between observed responses and those predicted by the geometric mean scaling model. The distribution of correlations is summarized in Table 4 as a function of matrix size. The last row in the table shows the percent of correlations that are significantly different from zero at each matrix size. It can be seen that the model reproduces the data to a reasonably good degree. For example, at matrix size 8, the model accounts for at least 62% of the response variance (0.79^2) in 50% of the cases, and for at least 41% (0.64^2) of the response variance in 75% of the cases.

Insert Table 4 about here

One may ask whether subjects judged some expressions with more internal consistency than others, so that the scaling model provided a better fit in those cases. The top part of Table 5 shows the mean over subjects of the linear correlation between observed and predicted responses for the geometric mean model separately for each expression. The Session 2 and Session 3 presentations of <u>possible</u> are combined, because they were not different. On 1 out of 20 scalings for <u>likely</u> and 5 out of 40 for <u>possible</u>, the resulting scale values were all equal. This could indicate either that the matrix contained only noise or that all the presented probabilities were equally well described by the term. Therefore, the associated

zero correlations between observed and predicted responses were not included in the means in Table 5. Note that tossup is fit considerably better than the other expressions on the average, but that otherwise there are no substantial differences among the terms.

Insert Table 5 about here

Concluding that the measurement models do a satisfactory job of scaling the degree to which a term better describes one probability than another, we now turn to the scale values to consider how reasonable they are as membership functions. For this purpose the derived values from each matrix were normalized by multiplication by a suitable constant so that the maximum value equaled 1. The values were plotted separately for each subject and each expression as a function of the spinner probabilities of landing on white. We will use the term membership function for the resulting graphs. Figure 5 illustrates the membership functions from three different subjects to show the range of results obtained

---

Insert Figure 5 about here

Subject 1, at the top of the figure, has monotonic membership functions with the exception of that for tossup. The remaining terms each span a range of probabilities, and the probability best described by each term is at one end of the range. Because subject 1 set the upper and lower probability limits for tossup equal to each other, its membership function is a point.

Subject 23, at the bottom of the figure, has membership

functions that tend to be single peaked. Thus, for this subject each expression spans a range of probabilities and the probability best described by that expression is somewhere in the center of the range.

Subject 6, in the middle of the figure, has both kinds of membership functions. This subject also illustrates functions that are not quite as well behaved, having two or even three peaks.

Recall that the functions for each expression were arbitrarily adjusted to have a maximum of 1, so that comparisons of ordinate values over terms is not meaningful. Also, ordinate values do not extend quite to zero, because the method of selecting probabilities based on Part 1 judgments purposely omitted probabilities with such membership values.

Table 6 summarizes the types of membership functions found over all subjects. The various functions can be characterized as either point, flat, monotonic increasing, monotonic decreasing, single peaked, or as having two, thre, or four peaks. The double peaked functions were further (subjectively) subdivided according to whether the second peak was very minor or not. The point, flat, monotonic, and single peaked functions might all be considered reasonable, in terms of the supposed underlying semantics, whereas the others cannot easily be so considered. Overall, 67% of the functions were reasonable by this criterion. If the double peaked functions in which one peak is minor are also included, then about 75% of the functions are reasonable and interpretable.

Insert Table 6 about here

If the multipeaked functions represent matrices that contain more rresponse error, then one would expect those matrices to have been less well fit by the scaling models. This indeed turned out to be the case. The 33% of the matrices leading to multipeaked functions have a mean goodness of fit correlation of 0.67 (S.D. = 0.16), while the 67% of the matrices leading to reasonable membership functions have a mean goodness of fit correlation of 0.83 (S.D. = 0.12).

The bottom part of Table 5 shows the percents of types of membership functions obtained for each term. For these purposes it was assumed that the multipeaked functions contained noise, and they were classified in with the flat, monotone increasing, single peaked or monotone decreasing functions as appropriate. It can be seen first that there was no expression for which all subjects had the same shape function. Second, terms closer to the extremes tended to have more monotonic than single peaked functions, whereas terms near the middle of the probability range tended to have more single peaked than monotonic functions. Tossup and almost impossible had point meanings for a few people. Finally, all forms of functions except point were obtained for possible.

However, even membership functions of the same type for a term did not look the same over subjects. The three expressions for which one might expect the most agreement on meaning are almost impossible, almost certain, and tossup. Their membership functions from all subjects are shown in Figure 6. Five subjects have point functions for tossup, two have single peaked functions that look different from the others, and the remaining 13 subjects show very similar functions. Almost impossible and almost certain

show considerable variability over subjects.

Insert Figure 6 about here

Expressions that are not near the anchor points of 0, 0.5, and 1 might be expected to show even greater individual differences, and they do. As one example, the membership functions for the word doubtful are shown in Figure 7. For purposes of clarity only, the monotonic functions are shown on the top half of the figure and the single peaked functions are shown on the bottom half. Note that some functions cover a large range and some a much smaller one. The peaks of the functions range from probability values close to zero to approximately 0.17. Analogous results hold for the other terms as well.

Insert Figure 7 about here

Part 3. The number of pairs of expressions and the number of probabilities per pair that a subject judged depended on the upper and lower limits set in Part 1. Recall that subjects always judged fewer pairs in Session 2 than in Session 3, because the latter included pairs with one member from each session as well as with both members from Session 3. Combining over both sessions, the number of pairs judged per subject ranged from 1 to 18 with a mean of 11.5 and a standard deviation of 4.4.

The number of probabilities judged per pair of expressions ranged from 1 to 8, except that, as already indicated, up to 16

probabilities were judged with pairs including possible in Session 3. Combining over sessions, the mean number of probabilities judged per pair was 8.4 with a standard deviation of 1.4.

The only analysis undertaken involved using Equation (7) to predict Part 3 judgments from membership function values derived in Part 2. Because the same probabilities were not generally presented with a term in the two parts, membership function values for probabilities used in Part 3 were estimated by linearly interpolating between the values derived for the two adjacent probabilities that were employed in Part 2. The ratios of the estimated values were then used in Equation (7) to predict the judgments, R, converted to ratios of distances, $S = R/(1-R)$, where, as before, $R = 0$ and 1 were converted to 0.0156 and 0.9844, respectively.

If $\beta_j, \beta_i = 1$ for all W_i and W_j in Equation (7), then within a pair of phrases the ratio of distances should be a linear function of the ratio of membership function values. This prediction was evaluated by means of a simple linear corrrelation pooled over phrase pairs for each subject in order to increase power. By pooling over expression pairs, the number of observations per correlation ranged from 7 to 146 over subjects (Mean = 83.2, S.D. = 37.7). The mean pooled correlation over subjects was 0.37, with a standard deviation of 0.23. Thirteen of the 20 correlations were significantly different from zero at $p < 0.005$, and two others were at $p < 0.10$.

Context effects. The membership functions for an expression are so variable over subjects that a powerful test of the effects of the different contexts for the two groups (cf. Table 1) is not

possible. One test of a context effect on the expressions in common to the two groups involves assigning three probabilities for each phrase to each subject. These values are the lower and the upper probability assigned by the subject in Part 1, and the "best" probability estimated from Part 2. The best probability for each term for each subject is the value that has a derived membership function of 1. For each term, a multivariate analysis of variance was then performed using the three probabilities transformed according to the equation $q = \ln(p/(1-p))$ to test for the difference between the two groups. The analysis was performed on q rather than on p to avoid the problems associated with a bounded scale. The multivariate tests showed no significant difference between the groups for any of the phrases.

Discussion

The results are quite encouraging overall, although in hindsight some design features were problematic. We will discuss the positive aspects first.

Part 1 provides the sole point of comparison between this study and others that have used a more traditional method to assess the meanings of probability phrases. The usual finding when subjects are asked to give numerical equivlents to probability phrases is considerable between-subject variability that is inversely related to distance from the center of the scale. This is precisely the pattern we obtained for the judgments of upper and lower probability limits.

The data of primary interest, of course, are from Part 2. Despite the lack of good inferential statistics, it seems

justifiable to say that the weak monotonicity axiom was well satisfied in the vast majority of cases. This, in conjunction with the fact that the other necessary conditions were forced to be satisfied by the data collection procedure, provided justification for applying the metric scaling models to the data. The scaling models fit well, accounting on the average for about 56% of the variance in the observed judgments without fitting a single free parameter. Clearly, nonmetric scaling procedures or procedures involving the estimation of free parameters would have done even better. Nevertheless, the derived scale values were generally of reasonable shape, and predicted the Part 3 responses to a relatively high degree. Thus, it appears justifiable to conclude that subjects can compare degrees of membership in a way that leads to consistent, meaningful and interpretable scaling of vagueness according to either a ratio or a difference model. However, it must be emphasized that nothing in the data allows us to determine whether subjects are more likely judging ratios or differences. (In a different situation, Birnbaum, e.g., 1980, concludes that subjects are judging differences not ratios, but to apply his empirical procedures here would tax even the most willing of subjects.)

Another conclusion that can be drawn from the present study is that even in this context, where the probabilities are well defined, there are large individual differences in the vague meanings of probability phrases. One might expect the potential for confusion to be greater in more natural contexts where the uncertainty itself is imprecise.

We allowed unique probability values to be associated with each expression over subjects because we expected considerable individual

differences, and because we were uncertain as to what probability range would be appropriate for a large group of people. However, there were problems that emerged as a result of individualizing the stimuli for each subject. First of all, all the trials in Parts 2 and 3 depended on a single determination of an upper and lower probability per term. If a subject made an error in Part 1 by setting a limit too high or too low, that error affected all the subsequent results. Note in Figure 6, for example, the two single peaked functions for tosssup that are different from the others. One would expect the derived membership function to extend closer to zero. If in Part 1 those two subjects had provided lesser lower bounds and greater upper bounds, then a larger range of probabilities would ave been presented to them in Part 2 and presumably more complete functions would have been derived. Similarly, note in Figure 5 Subject 1's membership functions for possible in Session 2 and Session 3. In both cases the subject gave zero as the lower limit for possible, but 0.5 was given as the upper limit in Session 2, whereas 1 was given as the upper limit in Session 3. The two functions look different because the maximum was forced to equal 1 in both cases.

As a further result of the strong reliance on the Part 1 judgments, there was no good way to evalute the stability over time of the membership functions for possible. This is because subjects tended to set different upper and lower probability limits for this term in the two sessions, resulting in different pair-comparisons. We originally selected possible on the assumption that it would cover the broadest probability range and therefore provide the most

sensitive test. In retrospect, the semantics of possible are very complex and it is possible that subjejcts attributed different meanings to the phrase in the two sessions.

There are yet two more consequences to having determined the Part 2 and 3 stimuli uniquely for each subject and for each part. One is that it was difficult to compare a phrase's membership functions over subjects. Membership functions that differ in shape are obviously distinct. However, two functions of the same shape may have distinct values at a given point, only due to the particular probabilities presented to the subject.

The other consequence is that predictions from Part 2 to Part 3 were weakened because it was necessary to base them on linear interpolations. Predictions would have been much more direct had they involved the same probabilities appearing with the phrase in Parts 2 and 3.

Finally, subjects did not report the task to be as difficult as we had originally envisioned it would be. Thus it might not have been necessary to have presented each probability pair only once for an expression in Part 2 and each expression pair only once for a probability in Part 3. Had each combination been presented at least twice (with the left-right ordering of the pairs reversed in half the trials) it would have been possible to have checked the sign reversal axiom, to have obtained a more thorough test of weak monotonicity (because more cells would have been involved), and to have determined empirically whether the response matrix was reciprocal. A second experiment was performed to correct these definciencies.

Experiment 2

This study had the same purposes as Experiment 1, but was designed to eliminate its problems. Specifically, the experiment was designed to better evaluate the stability of the membership functions over time, to more precisely determine the shape of the membership functions and to compare them over subjects, to obtain better tests of the algebraic difference structure axioms, to check reciprocity, and to better test the predictions from Part 2 to Part 3.

Method

Subjects. The four subjects in each group from Experiment 1 who had the highest mean goodness of fit corrrelations for the geometric mean scaling model were invited to take part in this study. Each was promised $15 for two sessions of approximately an hour and a quarter each.

Procedure. There was no Part 1. (However, for the sake of continuity, we will continue to denote the other two parts of the sessions as Part 2 and Part 3, respectively.) Rather, probabilities were selected on the basis of results from Experiment 1 and the same values were used for all subjects.

Insert Table 7 about here

Table 7 shows the expressions and associated probability values that were employed. Subjects in Group 1 judged doubtful, good chance, tossup, and probable, while those in Group 2 judged improbable, good chance, tossup and likely. Part 2 employed all pairs of the seven probabilities indicated for each term in the

table. Part 3 employed all pairs of terms for each indicated probability. Full left side by right side factorial designs were run within each session. That is to say, in Part 2 each distinct pair of probabilities was presented twice with each expression in each session, once in one left-right orientation and once in the reverse orientation. Similarly, in Part 3 each expression pair was presented twice with each probability in each session, once in each orientation. Within each part, presentation order was random, not blocked.

The response procedure was also changed, so that subjects used a joy stick to move the arrow on the response line. When the arrow was located in the desired position, the subject registered that response by pressing a button on the joy stick assembly. Whereas in the previous experiment the arrow could be located at any one of 17 discrete locations, the response line was essentially continuous in this study, limited only by the resolution of the screen.

Each subject was run on the full design within each of two sessions with approximately two days intervening.

Results

Reliability. Because Session 2 wa a replicate of Session 1, linear correlations can be used to assess reliability. Considering the response line to run from 1 to 0, Session 1 and 2 responses were correlated separately for Parts 2 and 3. The results are shown by subject and averaged over subjects in Table 8. In this and subsequent tables, marginal mean correlations are based on Fisher's r to z transformation. All subjects demonstrated quite high reliability, with Subject 9 showing the lowest Part 2 correlation

and Subject 14 showing the lowest Part 3 correlation. Considering this result, most subsequent analyses were done over the two sessions combined.

Insert Table 8 about here

Weak monotonicity. Utilizing mean responses over the two sessions, this axiom was checked in the same manner as Experiment 1. However, since the full $\underline{P} \times \underline{P}$ matrix was responded to for each phrase, $[\underline{n}!/(\underline{n}-3)!]^2 - [\underline{n}!/(\underline{n}-3)!]$ subsets of cells are available for test in each $\underline{P} \times \underline{P}$ matrix. For n = 7, a total of 43,890 subsets of cells can be tested for each phrase.

Table 9 shows mean satisfaction indices by subjects and by terms. It can be seen that the axiom is extremely well satisfied for all subjects, and that satisfaction is somewhat less for the terms doubtful and improbable than for the others.

Insert Table 9 about here

Sign reversal and reciprocity. If for a given $\underline{P} \times \underline{P}$ matrix the entry in cell $(\underline{p}_i, \underline{p}_j)$ is the complement of that in cell $(\underline{p}_j, \underline{p}_i)$, then the axiom of sign reversal is satisfied. In addition, the matrix for ratio scaling obtained by the transformation in Equation (4) will be reciprocal. An evaluation of complementarity is obtained by calculating the correlation for responses in cell $(\underline{p}_i, \underline{p}_j)$ as a function of those in cell $(\underline{p}_j, \underline{p}_i)$, as well as by fitting a linear structural model to these values (Isaac, 1970). Ideally, the correlation and the slope of the best fitting line will

both be -1. A linear structural model differs from a regression model in that it allows random error in both coordinates, not in just one. These analyses were applied to the response matrices for both Part 2 and Part 3. Mean slopes and pooled correlations for each subject are shown in Table 10, where it can be seen that the slopes and the correlations are very close to -1 for all subjects.

Insert Table 10 about here

Ratio scaling and membership functions. Part 2 responses were transformed according to Equation (4) (setting R = 0 and 1 equal to 0.004 and 0.9996, respectively), and the geometric mean scaling model was applied to them. Goodness of fit correlations are shown in Table 11 separately for each subject and phrase, but combined over sessions. It can be seen that goodness of fit is excellent, with the lowest correlation being 0.81 for likely for Subject 16.

Insert Table 11 about here

Normalizing the scale values from the separate matrices to have a maximum of 1 and plotting them as a function of the probabilities demonstrates that the Session 1 and Session 2 membership functions for each subject are quite similar, as would be expected given the high reliability. For illustration, the Session 1 and Session 2 membership functions are shown for two subjects in Figure 8. Subjects 4 and 14 were selected as having membership functions that appeared the most and the least similar, respectively. Even for

Subject 14 the membership functions are rather stable over the two sessions.

Insert Figure 8 about here

The membership functions obtained by applying the geometric mean model to the mean responses of Sessions 1 and 2 are shown in Figures 9 and 10, respectively, with a separate panel for each subject. As in Experiment 1, all subjects demonstrate similar membership functions for the word tossup. The functions for doubtful are also quite similar in shape, but those for the remaining expressions show remarkable differences over subjects.

Insert Figures 9 and 10 about here

The impressions about the membership functions obtained from visual inspection regarding within-subject stability and between-subject differences is substantiated by suitable statistical analysis of the underlying responses. For each phrase, there is a subject x session x probability pair factorial design with repeated measures over the last two factors. The number of observations per cell can be doubled by combining over reciprocal probability pairs, i.e., combining response $R_W(i,j)$ from cell (p_i,p_j) with response $(1-R_W(i,j))$ from cell (p_j,p_i). An 8 x 2 x 21, subject x sessions x probability pair, analysis of variance, assuming a randomized block design (Kirk, 1982, p. 441) was performed on the responses for good chance and tossup, and an identical 4 x 2 x 21 analysis of variance was performed on the responses for doubtful, improbable,

likely, and probable. In all cases there was an unsurprising, highly significant effect due to probability pairs (all $p < 0.0001$). Neither subject nor session was significant for tossup ($p = 0.998$ and 0.809, respectively), indicating that responses were stable over sessions and subjects. However, for all other terms there was a highly significant effect due to subjects ($p < 0.0001$ in all cases, except $p < 0.0002$ for improbable), reflecting the substantial differences in membership functions. For improbable, likely, and probable, there were no session effects ($p = 0.230$, 0.662, and 0.189, respectively), while there were small but significant session effects for doubtful and good chance ($p = 0.039$ and 0.024, respectively). Close inspection of the derived membership functions in the latter two cases suggested a slight sharpening of the functions for doubtful, and a slight broadening of the functions for good chance, from Session 1 to Session 2, as can be noted in Figure 8.

Predicting Part 3 responses. Note in Table 9 that probability values of 0.40, 0.45, and 0.50 were each associated with all four terms, whereas the remaining probabilities that appeared in Part 3 were associated with only two terms each. Thus, predictions are possible for trials that included 0.40, 0.45, and 0.50 only.

For each of these values, there was a left side by right side, term by term, factorial design, except of course omitting the diagonal cells. Scale values derived in Part 2 were used in Equation (7) to predict Part 3 responses transformed to a ratio of distances and combined over reciprocal cells for increased stability. On the assumption that all constants in the equation equal 1, the

prediction is evaluated by calculating the linear corrrelation between predicted and observed values at each of the three probabilities. Results are shown in Table 12, where it can be seen that the prediction is quite well sustained for all the subjects except 16 and 20.

Insert Table 12 about here

Finally, the factorial design at each of the three probabilities allows a geometric mean ratio scaling of the response matrices using equations analogous to Equations (4) - (6). As shown in Equation (8), the resulting membership function values, $\mu_p(\underline{W})$, should be a power function of those derived in Part 2, $\mu_W(\underline{p})$, if they both represent the same vagueness construct.

Power functions were fit to the scatter plot of $\mu_p(W)$ vs. $\mu_W(p)$ for each subject, and were assessed by means of F-ratios. The F-ratios ranged from 43.2 to 7,461 over subjects, with a median value of 143. Although inferential statistics are not appropriate (because the data points are not independent), it is clear descriptively that the functions fit very well. Those for Subjects 8 and 9, selected as a middling and the worst fit, and shown in Figure 11.

Insert Figure 11 about here

Discussion

This experiment seems to have overcome the problems of the first while substantiating its results. Because subjects were

selected for inclusion in this study on the basis of their scaling results in Experiment 1, it is perhaps not surprising that the algebraic difference structure axioms were well satisfied and that the geometric mean ratio scaling model described the judgments to a high degree in each case. However, it was necessary to obtain the good fits in order to properly test the other predictions.

The first notable result is that judgments were very stable over the two sessions, but differed considerably over subjects for all terms except tossup. As a consequence, membership functions for all the other terms varied widely and reliably over subjects. Tossup yielded similar single peaked functions for all eight subjects. Doubtful yielded different monotonic decreasing functions for the four subjects who judged it, while the remaining phrases resulted in both monotonic and single peaked functions. Furthermore, in these cases same shaped functions did not take on similar values, so that none of the remaining terms had precisely the same functions for any two subjects.

It is of interest to compare these membership functions to their counterparts in Experiment 1. Recall that the subjects in this experiment were also in the first one, and that they had judged the same expressions (among others) at that time. Because different probability values were used in the two studies, the only possible comparisons are in terms of membership function shapes. Of the 32 comparisons (8 subjects x 4 expressions each), derived membership functions were similar in shape in 25 cases. Of the remaining seven cases, six changed from point or monotonic to single peaked, and one changed from single peaked to monotonic decreasing. Subjects were

run in the two experiments at an interval of two to three months, so the similarity in 78% of the functions is striking.

On two grounds, it is reasonable to assume that the membership functions in this experiment in fact represented the vague meanings of the phrases to the subjects in this context. First, they all had sensible shapes. But of greater importance, they predicted independent judgments in Part 3 very well. Freed from the necessity of interpolation, ratios of membership function values derived in Part 2 correlated very highly with ratios of Part 3 responses converted to distances. In addition, membership function values independently derived from judgments in Parts 2 and 3 were related by a power function, as they were predicted to on the assumption that they were both measures of the same construct.

GENERAL DISCUSSION

Methodological issues

The present work is primarily methodological, but it has numerous substantive implications as well. Considering the methodological features first, we have demonstrated that in a specific context an individual's understanding of the vague meaning of a nonnumerical probability expression can be measured in a valid and reliable way. The measurement and scaling procedures employed here are based on a solid theoretical foundation, and overcome problems identified with other methods.

In particular, studies in which membership functions have been constructed from choice probabilities have been criticized as doing little more than relabeling measurement and sampling error as construct vagueness. Studies utilizing magnitude estimation procedures have addressed vagueness directly, but

frequently without a way to assess the meaningfulness or validity of the resulting scales. The procedures used in the present experiments avoided these problems. Subjects judged vagueness directly, in the sense of comparing degrees of membership of a stimulus in two ill-defined categories, within an experimental design that yielded three converging means for assessing the quality of the judgments.

First, conjoint-measurement provided the theoretical rationale for numerically scaling the judgments. Therefore, evaluation of the necessary conjoint-measurement axioms provided a means of evaluating the internal consistency of the judgments prior to numerical scaling. If the axioms had failed empirically, then we would have concluded that the subjects were not judging degrees of membership according to the difference or ratio rule that was to underlie the numerical scaling. Consequently, such scaling would have been inadmissible.

Because the axioms were generally well satisfied, the numerical scaling procedures were applied to the judgments. Goodness of fit measures, namely the correlations between observed and predicted responses, provided a second validity check. If the fits had been poor, then we would have concluded that the judgments were not represented well by the scales. We employed metric scaling procedures utilizing no free parameters, and, particularly in Experiment 2, achieved excellent fits. Had that not been the case, nonmetric methods with parameters fit to data could have been employed. Goodness of fit would have improved, but not necessarily to an acceptable level.

Although the two checks on the validity of the measurement procedures were passed, it is not necessary to conclude that subjects were judging the semantic vagueness of the terms. They could have been consistently judging some other quality instead. The third validity assessment, in the spirit of construct validity, was achieved by using the derived membership function values to predict independent judgments that were presumed to be based on the underlying vagueness dimension. The predictions were generally borne out, and consequently it appears justifiable to claim that the vague meanings of the terms were measured.

From the usual perspective of test theory, reliability is logically prior to validity and therefore must be established first. Judgments were reliable in Experiment 2 by the usual criteria, as were Part 1 upper and lower probabilities for possible in Experiment 1. However beyond the high test-retest correlations, the derived membership functions for each subject in Experiment 2 were very similar over the two sessions, and indeed generally reproduced the membership function shapes derived some 10 weeks earlier for corresponding terms in Experiment 1. This can be taken as further evidence that the subjects were judging an enduring property of the expressions.

On all the above grounds, we believe that the methodological aims of the study have been satisfied, and that we have established a means for validly measuring the vague meanings of nonnumerical probability expressions. Although we have not done so, there is no reason to think that the procedures could not be applied to other linguistic variables or vague categories as well.

Substantive issues

Numerous questions of substantive interest are raised by these results. They must be considered with regard to the intertwined practical issues of communication and theoretical issues of psycholinguistics and underlying cognitive processes.

Communication The claim that nonnumerical probability expressions convey vague uncertainties is clearly supported by the present data. It is noteworthy that such results were obtained despite the fact that the probabilistic events (spinner pointers landing on white) were exactly specified and easily judged numerically. As a check for the latter claim, three subjects subsequently provided numerical judgments of these probabilities with essentially no error. (Also see the paper by Wallsten, 1971, in which subjects gave virtually errorless probability estimates of physical spinners.) Thus the vagueness can be attributed to the verbal expressions, and not to the perceived uncertainty.

Of course, it is just when the uncertainty itself is ill defined that nonnumerical expressions are normally used. We expect that individual differences in understanding these expressions would be even greater in such ill defined situations than in the present context. Alternatively, it might be argued that the large individual differences emerged because each person developed his or her own strategy for coping with the unnatural task of using nonnumerical probability expressions in a situation involving precise probabilities. Consequently, individual differences would be less in more natural situations. The claim

strikes us as unlikely, but it cannot be dismissed at this point. However, the methodology can be extended easily to ill defined uncertainties where the competing claims can be investigated.

Assuming we are correct, it is difficult at this point to see any advantage in using an unrestricted set of verbal expressions rather than probability intervals or upper and lower probabilities (Wallsten, Budescu, and Forsyth, 1983) to express ill defined uncertainties. But it is conceivable that a subset of expressions can be determined by means of the present or another scaling technigue whose meanings are agreed upon by most people. There was good agreement on the meaning of tossup, and relatively less disagreement on the meanings of almost impossible and almost certain than of the other expressions, suggesting that there might exist a subset of agreed upon terms or phrases. In addition, it is possible that a commonly understood vocabulary naturally evolves among a group of experts all working in the same domain, although the results of Beyth-Marom (1982) suggest otherwise. The question of how best to communicate vague uncertainties from one person to another is still very much an open issue.

Shapes of Membership Functions. In the present domain, point, flat, monotonic increasing or decreasing, and single peaked are all reasonable shapes for membership functions. Nevertheless, our prior expectations, as well as those of at least some other authors, judging by illustrative functions that they have drawn (e.g., Watson, et al., 1979; Zimmer, 1973), was that the functions would be generally single peaked and that other shapes would be rare.

In contrast to our expectations, a slight majority of the membership functions were monotonic, 55% in Experiment 1 (from Table 8) and 56% in Experiment 2 (from Figures 9 and 10). With the exception of 6% of the functions in Experiment 1 which were point or flat, all the remaining were single peaked. It is of interest to speculate on the semantic differences implied by the single peaked versus monotonic functions, particularly since over the two experiments no expression was characterized by a single shape.

First, however, we cannot rule out the possibility that some aspect of the comparison procedure, such as inexperience with the task, artificially induced the monotonic functions. Weak evidence supporting this possibility is the fact that 6 of the 7 changes in membership functions from Experiment 1 to Experiment 2 were from monotonic to single peaked, and only 1 was in the reversed direction. However, in opposition to this possibility is the fact that Part 3 judgments were predicted equally well by both kinds of functions.

Assuming that the monotonic shapes are not artifactual, an aid to interpreting the differences between the two types of functions may come from considering functions empirically determined in other areas involving linguistic variables over numerical domains (Zadeh, 1975). Hersh and Caramazza (1976) considered the terms small and large along with the modifiers not, very, very very, not very, not very very, and sort of, as applied to squares of different areas. Hersh, Caramazza, and Brownell (1979) investigated the terms short and long alone and

with the modifiers very and sort of as applied to line lengths. Norwich and Turksen (1984) investigated tall, very tall, not tall, and short, and MacVicar-Whelan (1978) looked at tall and short, alone and in combination with very, all with regard to men's heights. Kuz'Min (1981) considered cold and warm along with numerous modifiers as applied to water temperature for swimming, as well as obsolete and up to date with and without modifiers as applie articles with regard to relevance. Finally, Zysno (1981) considered old and young, alone and with very applied to men's ages.

The studies used various empirical procedures, some of which we have taken issue with above, but generalizations do emerge. The majority of the membership functions were monotonic, with single peaked functions appearing especially under three conditions. First, they tended to appear with hedged expressions. A square is sort of large if it is neither too large nor too small. Second, single peaked functions sometimes occurred for an expression when a more extreme expression was also being considered. For example, if a square is very large, then for some people it is not large. Hersh and Caramazza (1976) referred to this interpretation of the expressions as linguistic, and distinguished it from the logical interpretation that occurred more frequently, in which a very large square is also large. Third, single peaked functions appeared for expressions, such as warm, that naturally occupy a midrange on a continuum.

The three conditions yielding single peaked functions may in fact be closely related to each other. Terms naturally occupy

the midrange of a continuum only when there are available expressions to describe the continuum on either side. Similarly, hedged expressions function as they do, because more extreme expressions are available (but not employed as the descriptors). And finally, the use of extreme terms may cause less extreme ones to have hedge-like qualities.

Applying these generalizations to our results, tossup refers to a probability roughly (or exactly) midway between 0 and 1, and the membership functions reflected that. All the other terms showed both kinds of functions. This suggests that some subjects considered specific expressions to be hedges, i.e., to refer to probabilities that are neither too large nor too small, and to be restricted in meaning by the availability of other terms. Other subjects considered the expressions in a non-hedged fashion, independently of other terms. As such, this explanation is no more than a restatement of our data, but it has interesting empirical implications.

First, the shape of the membership function for tossup should be unaffected by the other expressions under consideration. Second, more subjects should demonstrate single peaked functions for a term if it is considered in conjunction with itself modified by very, or in conjunction with obviously extreme terms (e.g., almost certain), than if it is considered alone. However, subjects demonstrating many monotonic functions may be more inclined to consider each expression in isolation, and therefore be less sensitive to the effects of such modifiers and extreme expressions than other subjects.

Another prediction is that subjects are more likely to provide single peaked functions when they are selecting an expression to apply to a situation, and therefore are considering alternative expressions, than when they hear or read the expression from others and are not considering alternatives. It might also be predicted that subjects who give evidence of many single peaked functions differentiate levels of uncertainty to a greater extent than do other subjects.

There may be semantic factors that determine whether a probability expression is interpreted in a hedged or non-hedged fashion. For example, perhaps the hedged interpretations are more common in political or human contexts while the non-hedged are in scientific or physical contexts. More generally, a single peaked function suggests that the uncertainty is being specified with both an upper and a lower bound, while a monotonic function suggests that it is being specified with only one of the two bounds. It is of interest to uncover the factors that determine each kind of processing of vague uncertainty.

Comparing membership values over expressions. In deriving the membership functions, the maximum membership value for each expression was arbitrarily set to one. For this reason, it is not meaningful to compare degrees of membership of probability values in different terms. However, such comparisons could be elicited directly so that relative heights of membership functions could be determined.

For example, consider the functions for Subject 14 in Experiment 2 (Figure 10) for good chance and tossup, which peak at 0.55 and 0.50, respectively. If we had asked the subject how

much better tossup described 0.50 than good chance described 0.55 (or conversely), we might have adjusted the two functions appropriately.

Similarly, it appears from the functions in Figures 9 and 10 that good chance and probable are synonomous for two subjects, as are good chance and likely for two others. This is not necessarily the case, since we cannot tell whether any of the subjects consider one expression to be uniformly better than the other over the range of probabilities used. In the same vein, when two functions cross, we cannot necessarily claim that the subject's judgment of the relative goodness of the two expressions also switches at that point.

It should be noted that the predictions of Part 3 from Part 2 responses did not require ratio comparisons of the sort being discussed here. The predictions were made respecting the levels of uniqueness determined by the algebraic difference structure axioms.

The question of relative membership values for different expressions is an interesting one. For example, some terms such as possible might have uniformly low membership function values, while others such as tossup or almost certain have high values for some probabilities. The relative heights of functions might be related to an individual's propensity to use the various expressions, and might change with different context factors.

Context effects. The meanings of nonnumerical probability phrases, even to an individual are almost assuredly not fixed over contexts. Many possible context effects have already been

discussed, but others should also be mentioned.

For example, Pepper and Prytulak (1974) have shown that the interpretations of relative quantifiers such as frequently or sometimes depend on the expected frequency of the event being described; Cohen, Dearnley, and Hansel (1958) have shown that the interpretations of quantifiers of amount such as some or several depend on the available quantity; and Wallsten, Fillenbaum, and Cox (in preparation) have shown that the interpretations of probability expressions depend on the base rate of the event in question. In addition, we may speculate that event importance and desirability also affect the meanings of probability expressions.

Zimmer (1984) has suggested that the interpretations of probability expressions vary over knowledge domains. The possibility was raised above that the shapes of membership functions may change with the area of discourse. Less dramatic, but equally interesting, would be changes in probability ranges covered by an expression, or changes in relative magnitudes of membership values, as a function of knowledge domain.

The procedures developed in the present experiments provide some insight into individuals' use of probability terms. More importantly however, they provide the means for investigating questions of the sort raised in the latter part of this discussion regarding how people form and communicate vague opinions.

References

Ballmer, T. T., & Pinkal, M. (Eds.) (1983). Approaching vagueness. Amsterdam: North-Holland.

Bass, B. M., Cascio, W. F.,& O'Connor. (1974). Magnitude estimation of expressions of frequency and amount. Journal of Applied Psychology, 59, 313-320.

Beyth-Marom, R. (1982). How probable is probable? Numerical translation of verbal probability expressions. Journal of Forecasting, 1, 257-269.

Birnbaum, M. H. (1980). Comparison of two theories of "ratio" and "difference" judgments. Journal of Experimental Psychology: General, 109, 304-319.

Budescu, D. V. (1984). Scaling binary comparison matrices: A comment on Narasimhan's proposal and other methods. Fuzzy Sets and Systems, 14, 187-192.

Budescu, D. V., & Wallsten, T. S. (in press). Consistency in interpretation of probabilistic phrases. Organizational Behavior and Human Decision Processes.

Budescu, D. V., Zwick, R., & Rapoport, A. (1985). A comparison of the analytic hierarchy process and the geometric mean procedure for ratio scaling (Report No. 172). Chapel Hill: University of North Carolina, L. L. Thurstone Psychometric Laboratory.

Cohen, J., Dearnley, E. J., and Hansel, C.E.M. (1958). A quantitative study of meaning. British Journal of Educational Psychology, 28, 141-148.

De Jong, P. (1984). A statistical approach to Saaty's scaling method for priorities. Journal of Mathematical Psychology, 28,

467-478.

Foley, B. J. (1959). The expression of certainty. American Journal of Psychology, 72, 614-615.

Hempel, C. G. (1939). Vagueness and logic. Philosophy of Science, 6, 163-180.

Hersh, H. M., & Caramazza, A. (1976). A fuzzy set approach to modifiers and vagueness in natural language. Journal of Experimental Psychology: General, 105, 254-276.

Hersh, H. M., Caramazza, A., & Brownell, H. H. (1979) Effects of context on fuzzy membership functions. In M. M. Gupta, R. K. Ragade, & R. R. Yager (Eds.), Advances in fuzzy set theory and application. Amsterdam: North-Holland, 1979.

Gaines, B. R., & Kohout, L. J. (1977). The fuzzy decade: A bibliography of fuzzy systems and closely related topics. International Journal of Man-Machine Studies, 9, 1-68.

Goguen, J. A. (1969). The logic of inexact concepts. Synthese, 19, 325-373.

Gulliksen, H. (1959). Mathematical solutions for psychological problems. American Scientist, 47, 178-201.

Isaac, P. D. (1970). Linear regression, structural relations, and measurement error. Psychological Bulletin, 74, 213-218.

Jensen, R. E. (1984). An alternative scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology, 28, 317-332.

Johnson, C. R., Bline, W. B., & Wang, T. Y. (1979). Right-left asymmetry in an eigenvector ranking procedure. Journal of Mathematical Psychology, 19, 61-64.

Johnson, E. M. (1973) Encoding of qualitative expressions of uncertainty (Technical Paper 250). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD-780 814)

Kirk, R. E. (1982). Experimental design (second edition). Belmont, CA: Brooks/Cole Publishing Co.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). Foundations of measurement. Vol. I. New York: Academic Press.

Kuz'min, V. B. (1981) A parametric approach to description of linguistic values of variables and hedges. Fuzzy Sets and Systems, 6, 27-41.

Labov, W. (1973). The boundaries of words and their meanings. In C.-J. N. Bailey & R. W. Shuy (Eds.), New ways of analyzing variation in English. Washington, DC: Georgetown University Press.

Lichtenstein, S. & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. Psychonomic Science, 9, 563-564.

MacVicar-Whelan, P. J. (1978). Fuzzy sets, the concept of height, and the hedge very. IEEE Transactions on Systems, Man, and Cybernetics, SMC-8, 507-511.

Miyamoto, J. M. (1983). An axiomatization of the ratio/difference representation. Journal of Mathematical Psychology, 27, 439-455.

National Research Council Governing Board Committee on the Assessment of Risk (1981). The handling of risk assessments in NRC Reports. Washington, DC: U.S. National Research Council.

Norwich, A. M., & Turksen, I. B. (1982). The fundamental measurement of fuzziness. In R. R. Yager (Ed.), Fuzzy set and

possibility theory. New York: Pergamon Press.

Norwich, A. M., & Turksen, I. B. (1984). A model for the measurement of membership and the consequences of its empirical implementation. Fuzzy Sets and Systems, 12, 1-25.

Oden, G. C. (1977a). Integration of fuzzy logical information. Journal of Experimental Psychology: Human Perception and Performance, 3, 565-575.

Oden, G. C. (1977b). Fuzziness in semantic memory: Choosing exemplars of subjective categories. Memory and Cognition, 5, 198-204.

Oden, G. C. (1981). A fuzzy propositional model of concept structure and use: A case study in object identification. In G. W. Lasker (Ed.), Applied systems research and cybernetics. Elmsford, NY: Pergamon Press.

Pepper, S., & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effects in the interpretations of quantitative expressions. Journal of Research in Personality, 8, 95-101.

Rubin, D. C. (1979). On measuring fuzziness: A comment on "A fuzzy set approach to modifiers and vagueness in natural language." Journal of Experimental Psychology: General, 108, 486-489.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology, 15, 234-281.

Saaty, T. L. (1980). The analytic hierarchy process. New York: McGraw-Hill.

Saaty, T. L., & Vargas, L. G. (1984). Inconsistency and rank preservation. Journal of Mathematical Psychology, 28, 205-214.

Simpson, R. H. (1944). The specific meanings of certain terms indicating differing degrees of frequency. Quarterly Journal of Speech, 30, 328-330.

Simpson, R. H. (1963). Stability in meanings for quantitative terms: A comparison over 20 years. Quarterly Journal of Speech, 49, 146-151.

Sjöberg, L. (1980). Similarity and correlation. In E. J. Lantermann & H. Feger (Eds.), Similarity and choice. Bern: Hans Huber Publishers.

Skala, H. J., Termini, S., & Trillas, E. (1984). Aspects of vagueness. Dordrecht: Reidel.

Torgerson, W. S. (1958). Theory and Methods of Scaling. New York: Wiley.

Wallsten, T. S. (1971). Subjectively expected utility theory and subjects' probability estimates: Use of measurement-free techniques. Journal of Experimental Psychology, 88, 31-40.

Wallsten, T. S., Budescu, D. V., & Forsyth, B. H. (1983). Stability and coherence of health experts' upper and lower subjective probabilities about dose-response curves. Organizational Behavior and Human Performance, 31, 277-302.

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (in preparation). Base rate effects on the interpretation of probability and frequency expressions.

Watson, S. R., Weiss, J. J., & Donnell, M. L. ((1979). Fuzzy decision analysis. IEEE Transactions on Systems, Man, and Cybernetics, SMC-9, 1-9.

Williams, C., & Crawford, G. (1980). Analysis of subjective judgment matrices (Report R-2572-AF). Santa Monica, CA: Rand

Corporation.

Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8, 338-353.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning (II). Information Sciences, 8, 301-357.

Zimmer, A. C. (1983). Verbal vs. numerical processing of subjective probabilities. In R. W. Scholz (Ed.), Decision making under uncertainty. Amsterdam: North-Holland Publishers.

Zimmer, A. C. (1984). A model for the interpretation of verbal predictions. International Journal of Man-Machine Studies, 20, 121-134.

Zysno, P. (1981). Modeling membership functions. In B. B. Rieger (Ed.), Empirical semantics. Bochum: Brockmeyer.

Authors' Notes

Please address reprint requests to Thomas S. Wallsten, L. L. Thurstone Psychometric Laboratory, Davie Hall 013A, University of North Carolina, Chapel Hill, NC, 27514. This research was supported by Contract MDA 903-83-K-0347 from the U.S. Army Research Institute for the Behavioral and Social Sciences to the L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill. The views, opinions, and findings contained in this paper are those of the authors and should not be construed as an official Department of the Army position, policy, or decision. Barbara Forsyth is now at the University of Michigan. We thank Samuel Fillenbaum for numerous helpful discussions throughout the course of the work, and James Cox, Brent Cohen, Samuel Fillenbaum, and Jaan Valsiner for comments on a previous draft of this paper.

Table 1

Terms Used in Experiment 1

	Group 1	Group 2
Session 2		Almost impossible
	Doubtful	Doubtful
	Possible	Possible
	Tossup	Tossup
	Likely	Likely
	Almost certain	
Session 3		Unlikely
	Improbable	Improbable
	Possible	Possible
	Good chance	Good chance
	Probable	

Table 2

Frequencies of Values of $\Delta p = p^* - p_*$ and n in Experiment 1

Δp	n	Frequency
$\Delta p = 0$	0	8
$0 < \Delta p < .02$	0	0
$.02 \leq \Delta p < .04$	1	0
$.04 \leq \Delta p < .06$	2	0
$.06 \leq \Delta p < .08$	3	5
$.08 \leq \Delta p < .10$	4	4
$.10 \leq \Delta p < .12$	5	5
$.12 \leq \Delta p < .14$	6	7
$.14 \leq \Delta p < .16$	7	7
$.16 \leq \Delta p \leq 1.0$	8	144

Table 3

Summary of Satisfaction Indices for Weak Monotonicity in Experiment 1

	Matrix Size				
	4	5	6	7	8
			Random Data		
Mean	53.5	57.1	57.3	52.1	52.4
S.D.	18.6	12.7	2.5	6.3	4.8
			Actual Data		
Number of matrices	4	5	7	7	144
25th percentile	83	78	77	76	75
50th percentile	83	90	82	83	82
75th percentile	92	93	90	92	89
Percent for which z>3.0	0	0	100	100	91

Note: The satisfaction index is the percent of submatrices for which the antecedent conditions are met that also satisfy the consequent condition.

Table 4

Summary of Linear Correlations Between Observed and Predicted Responses for the Geometric Mean Scaling Model in Experiment 1

	Matrix Size					
	3	4	5	6	7	8
Number of matrices	5	4	5	7	7	144
25th percentile	.94	.52	.59	.59	.54	.64
50th percentile	1.00	.82	.72	.79	.68	.79
75th percentile	1.00	.95	.80	.88	.82	.87
Percent for which $p<.01$	80	50	60	86	71	83

Table 5

Mean Goodness of Fit Correlations and Percents of Different Shapes of Membership Functions for Each Term in Experiment 1

	Almost Certain	Probable	Likely	Good Chance	Possible	Tossup	Unlikely	Improb-able	Doubtful	Almost Impos-sible
					Goodness of fit correlations					
	.76	.73	.75[a]	.73	.73[a]	.93	.70	.80	.76	.85
					Percent different membership function shapes					
Point						25				20
Flat			5		12					
Monotonic Increasing	90	60	45	45	12					
1 peak	10	40	50	45	58	75	50	20	25	20
Monotonic decreasing				10	18		50	80	75	60

[a]Excluding solutions with equal scale values; see text.

Table 6

Percents of Different Shapes of Membership Functions in Experiment 1

Shape	Percent
Point	4
Flat	2
Monotonic	30
1 Peak	31
2 Peaks	
1 minor	8
other	18
3 Peaks	7
4 Peaks	0.5

Table 7

Expressions and Probability Values Used in Experiment 2

	Probabilities (x 100)												
	5	20	30	40	45	47	50	53	55	60	75	85	95
Group 1													
Probable				X	X		X			X	X	X	X
Good chance			X	X	X		X		X		X		X
Tossup				X	X	X	X	X	X	X			
Doubtful	X	X	X	X	X	X	X						
Group 2													
Likely				X	X		X			X	X	X	X
Good chance			X	X	X		X		X		X		X
Tossup				X	X	X	X	X	X	X			
Improbable	X	X	X	X	X	X	X						

Table 8

Linear Correlations Between Sessions 1 and 2 Responses in Experiment 2

Subject	Part 2	Part 3
1	.89	.93
4	.93	.83
8	.96	.81
9	.75	.89
14	.78	.61
16	.97	.98
17	.89	.93
20	.88	.95
$\bar{r}$	.90	.90

Table 9

Satisfaction Indices for the Weak Monotonicity Axiom, in Experiment 2

Subject	Index	Phrase	Index
1	90	Probable	92
4	90	Likely	93
8	91	Good chance	92
9	86	Tossup	94
14	90	Improbable	83
16	91	Doubtful	79
17	90		
20	91		

Table 10

Mean Slopes from the Linear Structural Model and Pooled Correlations for Responses in Cell (j,i) as a Function of Cell (i,j), Over Terms and Sessions in Experiment 2

	Part 2		Part 3	
Subject	Slope	Correlation	Slope	Correlation
1	-1.00	-.92	-1.01	-.96
4	-1.00	-.93	-.97	-.97
8	-1.05	-.94	-.87	-.86
9	-.98	-.85	-.88	-.95
14	-.94	-.88	-1.08	-.85
16	-1.03	-.98	-.98	-.99
17	-1.00	-.88	-.91	-.96
20	-.92	-.92	-1.02	-.95
$\bar{X}$	-.99	-.92	-.97	-.95

Table 11

Mean Linear Correlations Between Observed and Predicted Responses for the Geometric Mean Model, Experiment 2

Subject	Probable	Likely	Good Chance	Tossup	Improbable	Doubtful	$\bar{r}$
1	.98		.88	.93		.97	.95
4	.97		.96	.96		.98	.97
8	.98		.99	.96		.98	.98
9	.78		.91	.83		.96	.89
14		.91	.89	.97	.97		.95
16		.81	.84	1.00	.97		.95
17		.98	.97	.88	.96		.96
20		.95	.92	.94	.98		.95
$\bar{r}$	.96	.93	.94	.95	.97	.97	.95

Table 12

Linear Correlations Between Observed and Predicted Part 3 Responses Transformed to Distance Ratios in Experiment 2

	Probability			
Subject	.40	.45	.50	$\bar{r}$
1	.71	.77	.99	.91
4	.85	.95	.95	.93
8	.87	.74	.89	.84
9	.73	.97	.85	.89
14	.60	.86	.98	.89
16	.43	.57	.88	.68
17	.93	.78	.82	.86
20	.76	.27	.55	.56
$\bar{r}$	.77	.82	.92	.85

Figure Captions

Figure 1. Hypothetical membership functions for two probability terms.

Figure 2. Sample experimental scenario.

Figure 3. Illustration of the weak monotonicity axiom

Figure 4. First, second, and third quartiles over subjects of the upper and lower probability limits for each phrase in Experiment 1.

Figure 5. Derived membership functions for three subjects in Experiment 1. The functions are coded as follows: AC = *almost certain*; AI = *almost impossible*; D = *doubtful*; GC = *good chance*; I = *improbable*; L = *likely*; Po = *possible*; Pr = *probable*; T = *tossup*; U = *unlikely*.

Figure 6. Membership functions from all subjects in Experiment 1 for *almost certain* (AC), *almost impossible* (AI), and *tossup* (T).

Figure 7. Membership functions from all subjects in Experiment 1 for *doubtful*.

Figure 8. Membership functions by session for two subjects in Experiment 2. The functions are coded as follows: D = *doubtful*; GC = *good chance*; I = *improbable*; L = *likely*; P = *probable*; T = *tossup*.

Figure 9. Membership functions for Subjects 1, 4, 8, and 9 in Experiment 2. The functions are coded as follows: D = *doubtful*; GC = *good chance*; Pr = *probable*; T = *tossup*.

Figure 10. Membership functions for Subjects 14, 16, 17, and 20 in Experiment 2. The functions are coded as follows: GC = *good chance*; I = *improbable*; L = *likely*; T = *tossup*.

Figure 11. Power functions fit to $\mu_p(W)$ as a function of $\mu_W(p)$ for two subjects in Experiment 2.

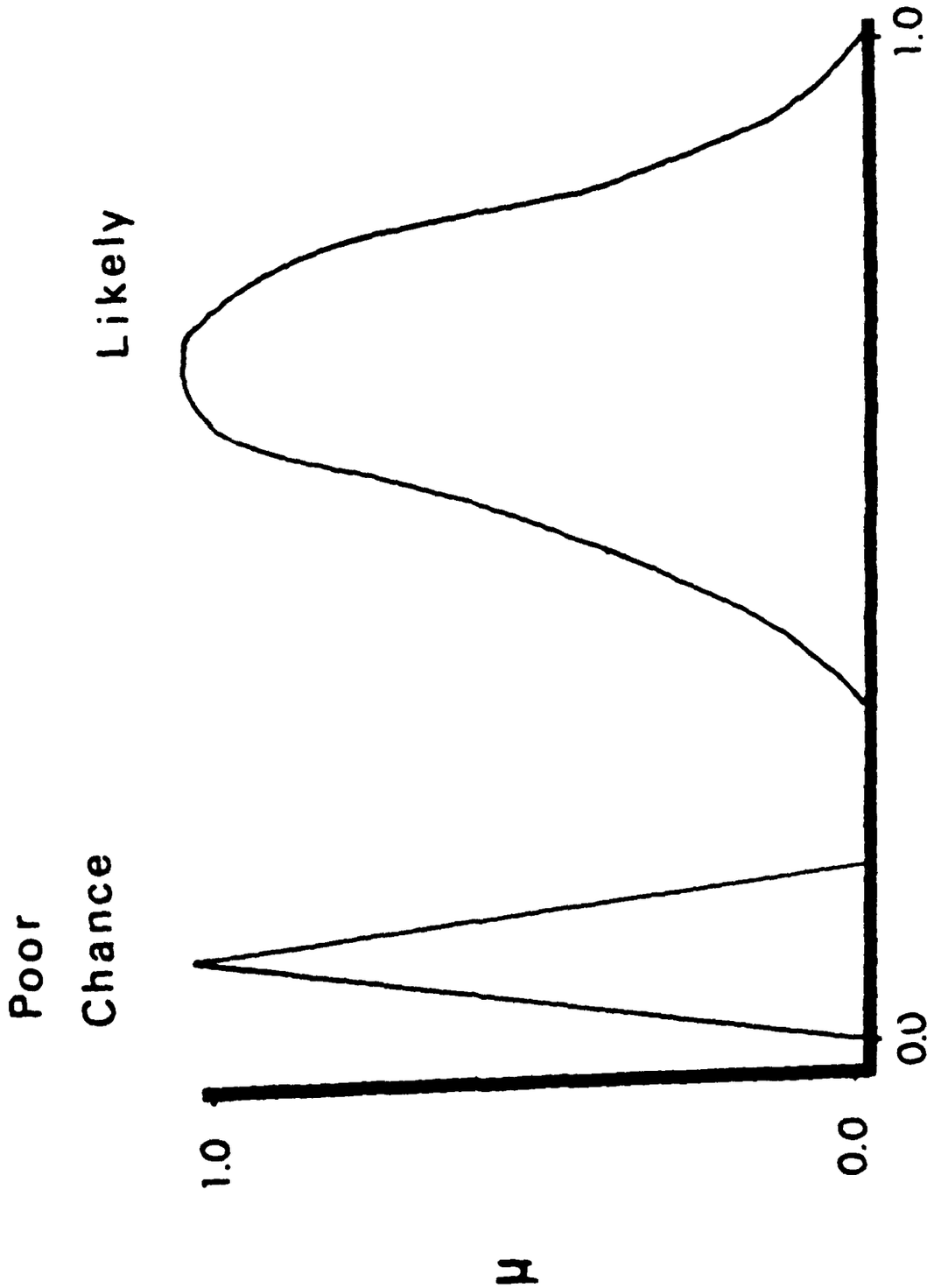

Fig. 1

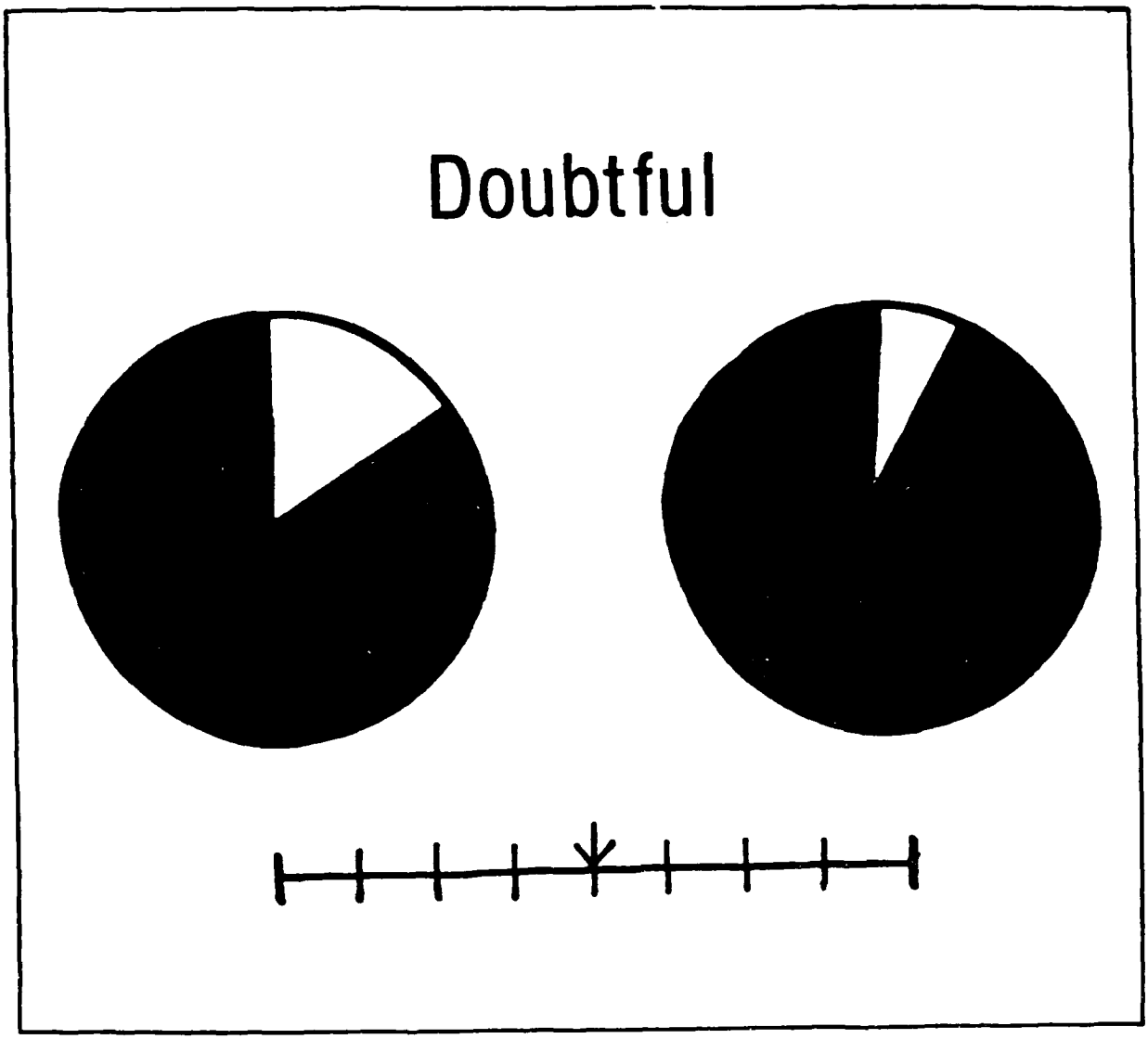

Fig. 2

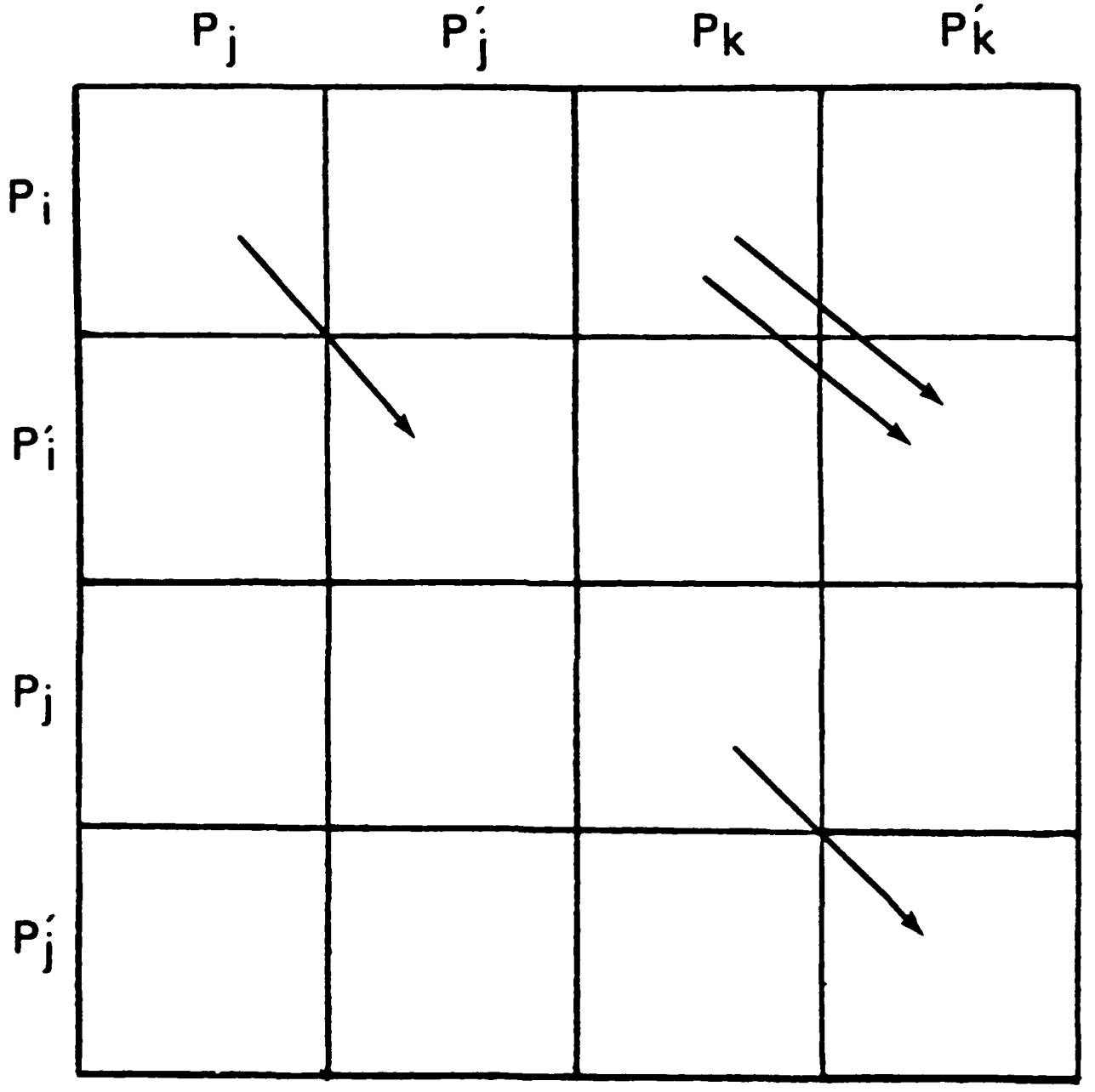

Fig. 3

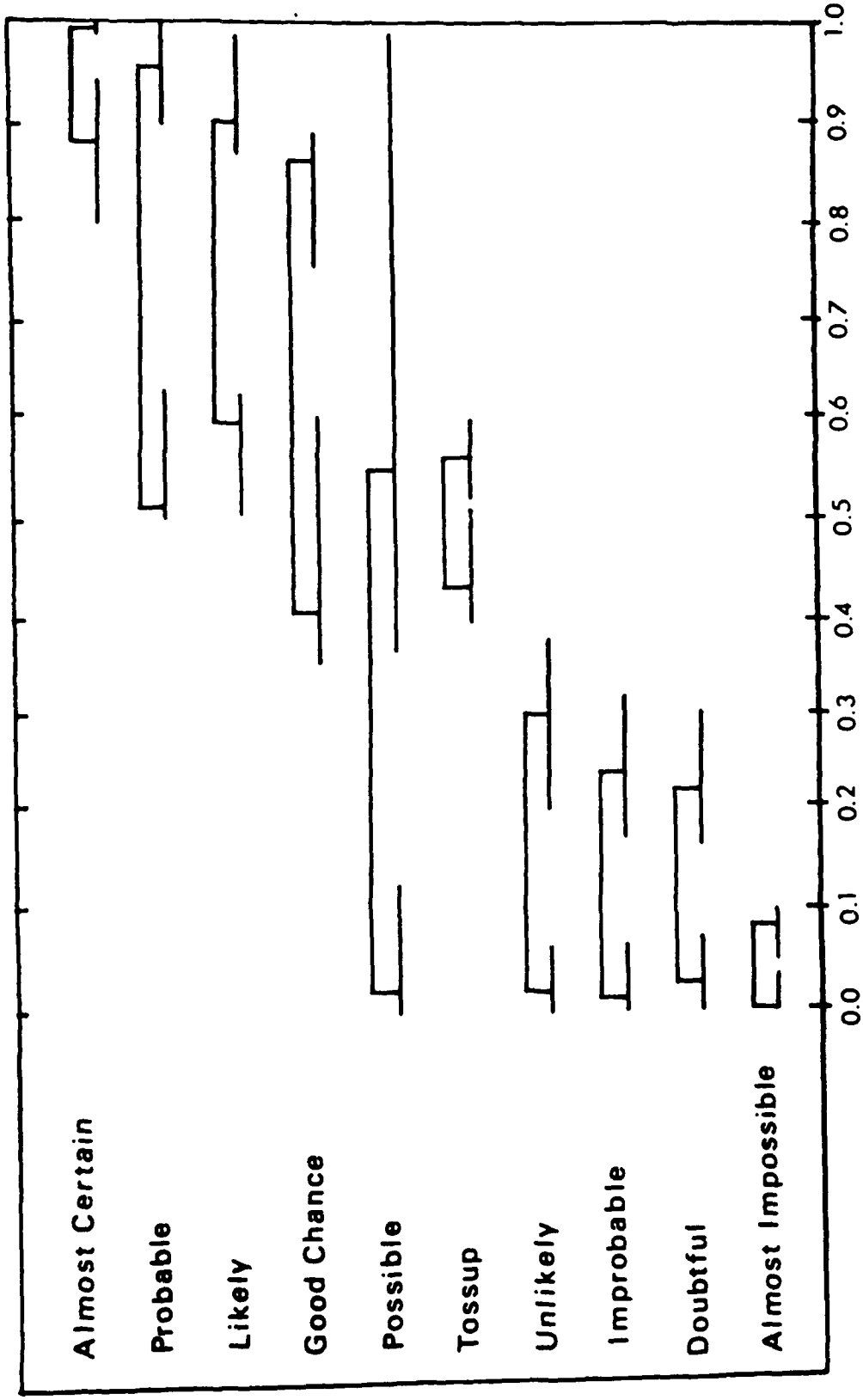
Almost Certain
Probable
Likely
Good Chance
Possible
Tossup
Unlikely
Improbable
Doubtful
Almost Impossible
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Probability

Fig. 4

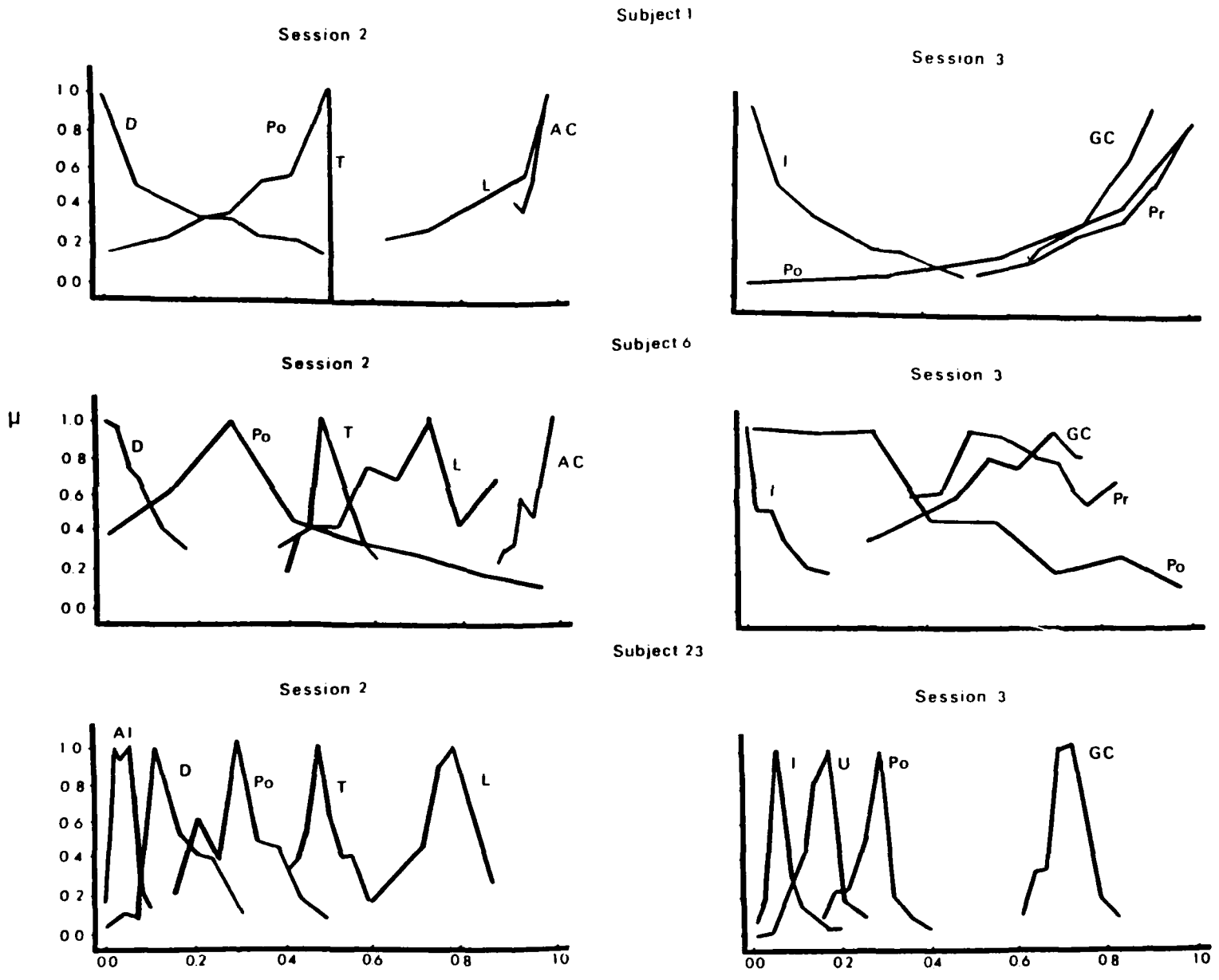
Subject 1
Session 2
Session 3
1.0
0.8
0.6
0.4
0.2
0.0
D
Po
T
L
AC
I
GC
Pr
Po
Subject 6
Session 2
Session 3
μ
D
Po
T
L
AC
I
GC
Pr
Po
Subject 23
Session 2
Session 3
AI
D
Po
T
L
I
U
Po
GC
0.0
0.2
0.4
0.6
0.8
1.0
Probability

Fig. 5

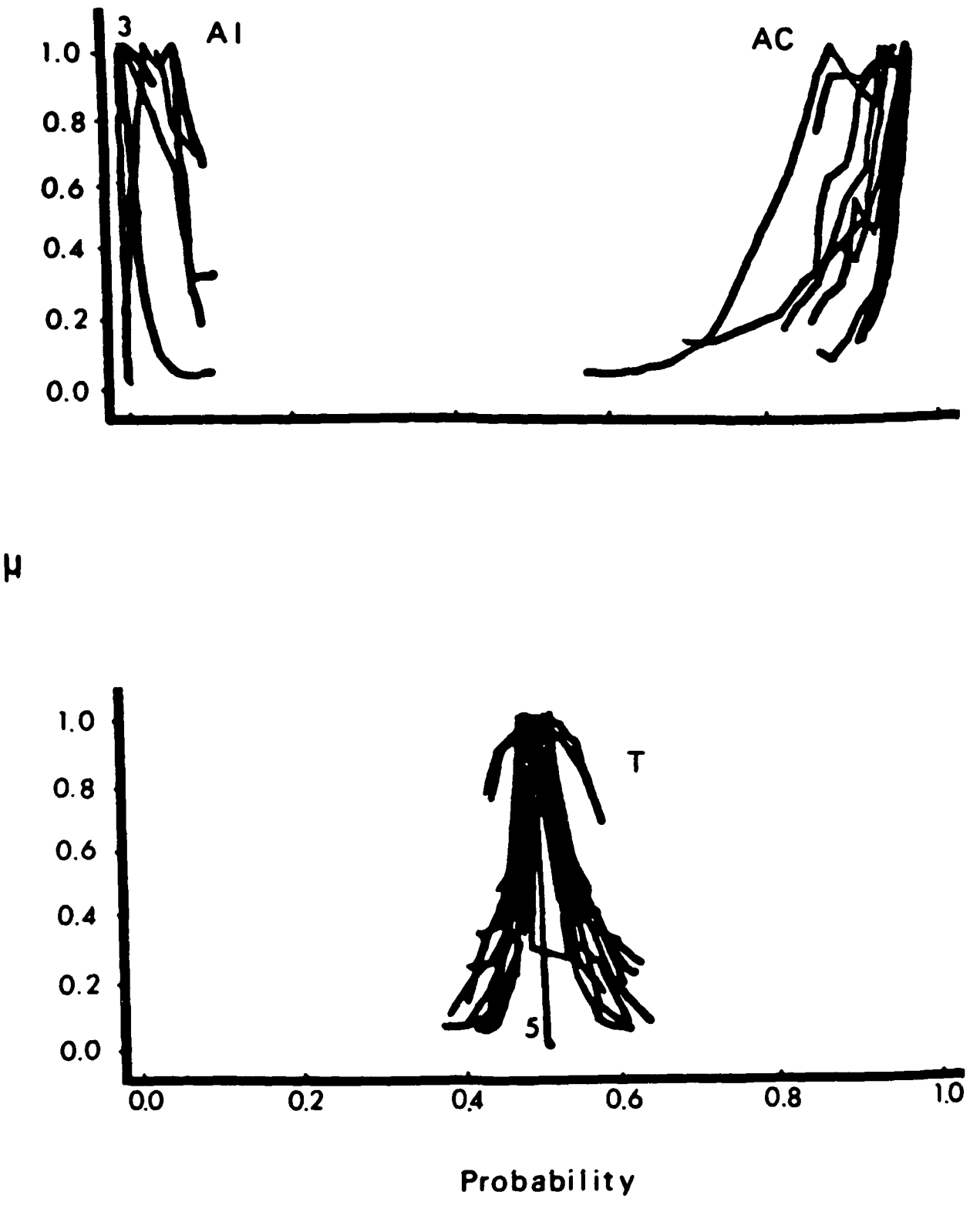

Fig. 6

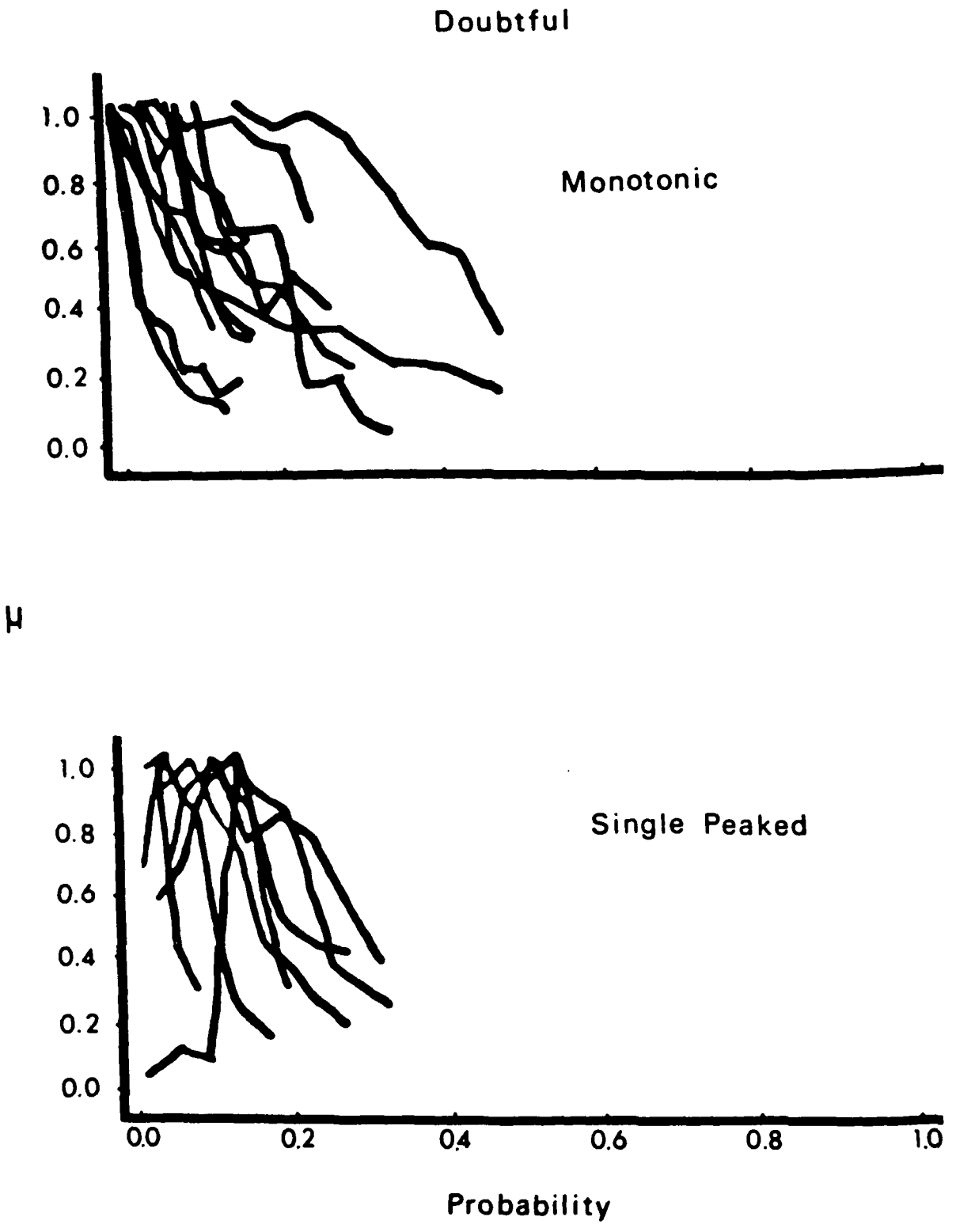
Doubtful
1.0
0.8
0.6
0.4
0.2
0.0
Monotonic
μ
1.0
0.8
0.6
0.4
0.2
0.0
Single Peaked
0.0
0.2
0.4
0.6
0.8
1.0
Probability

Fig. 7

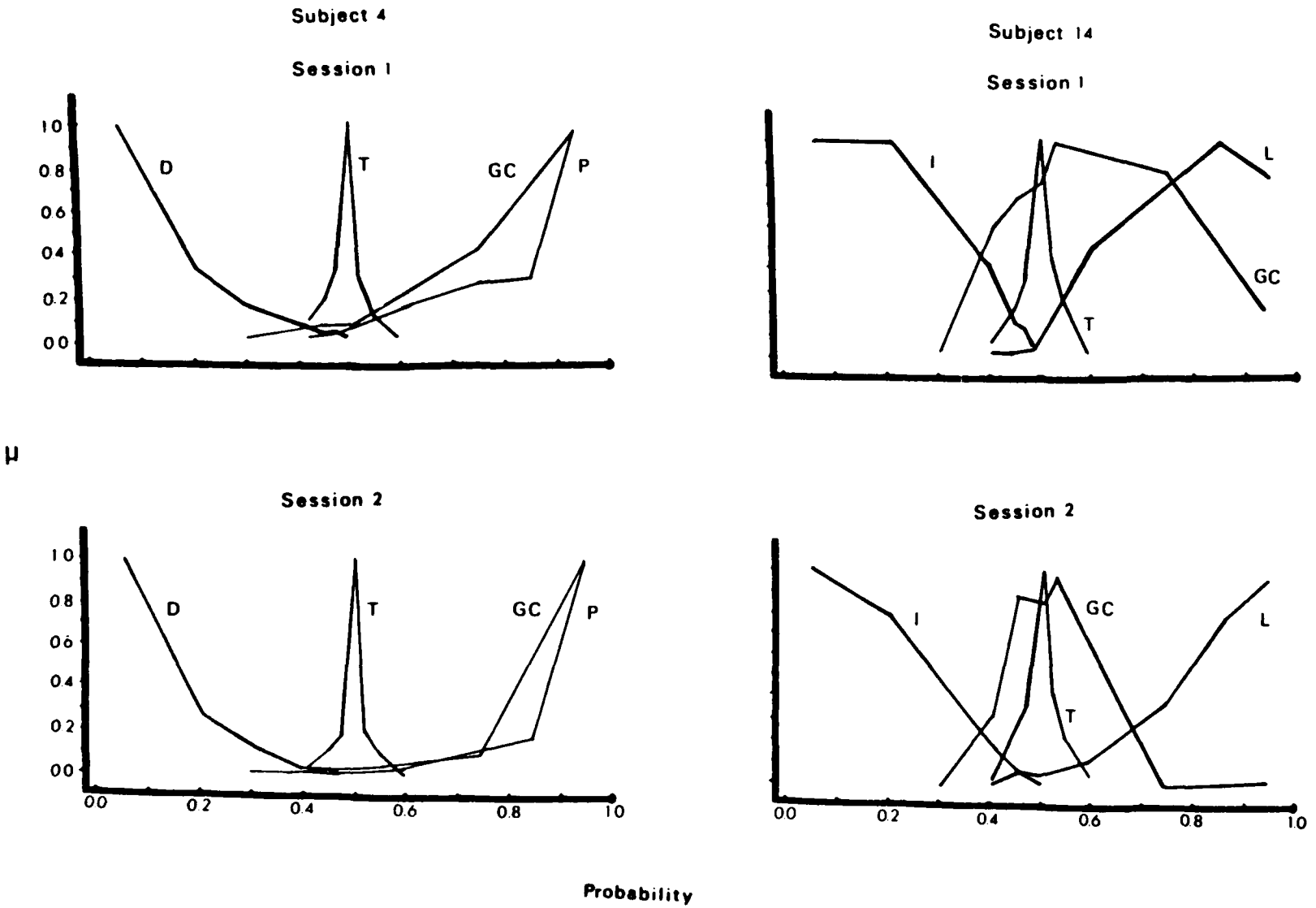
Subject 4
Session 1
D
T
GC
P
Subject 14
Session 1
I
L
GC
T
μ
Session 2
D
T
GC
P
Session 2
I
GC
L
T
0.0
0.2
0.4
0.6
0.8
1.0
Probability

Fig. 8

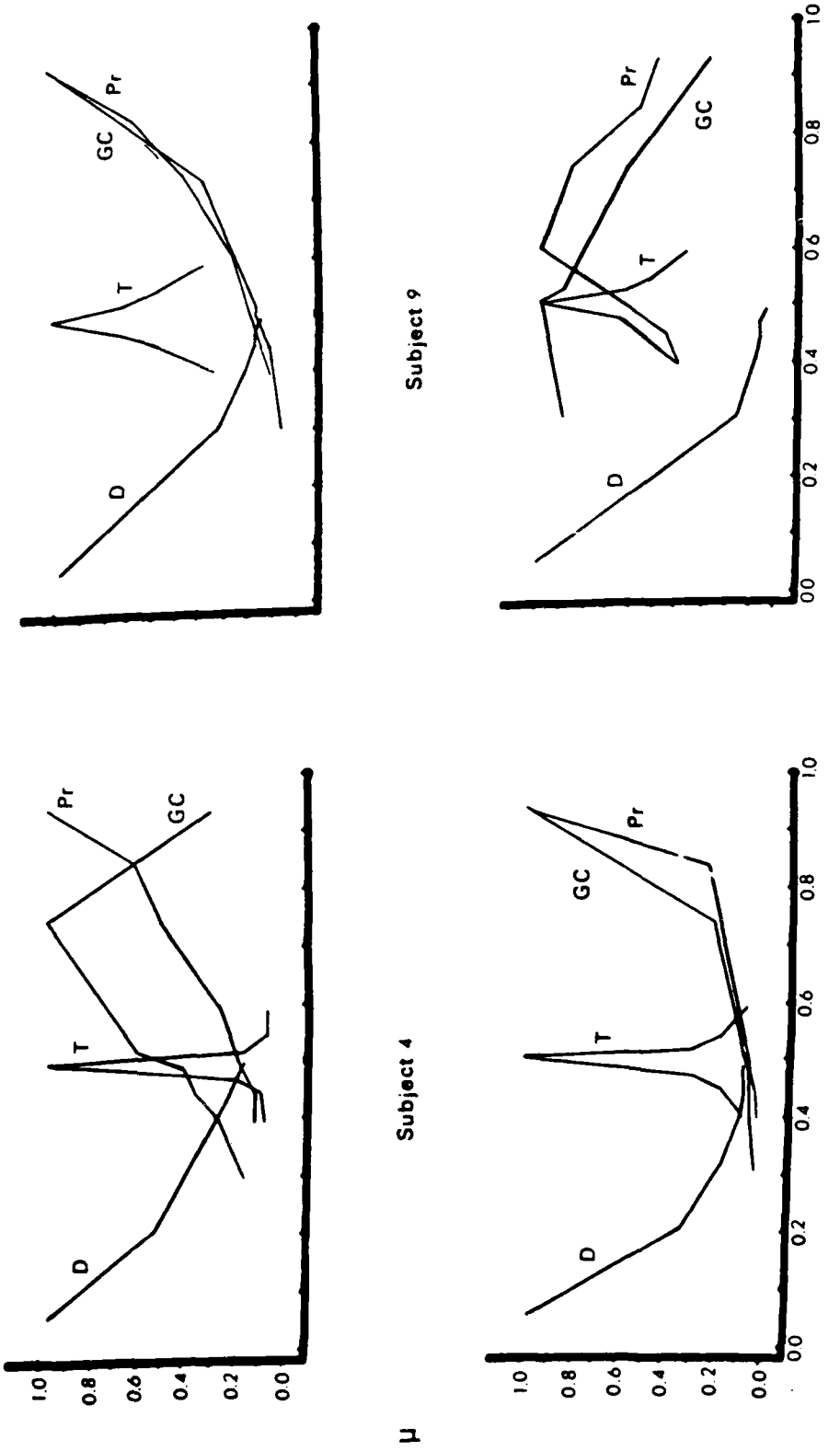
Subject 1
Subject 8
Subject 4
Subject 9
D
T
GC
Pr
μ
Probability
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 9

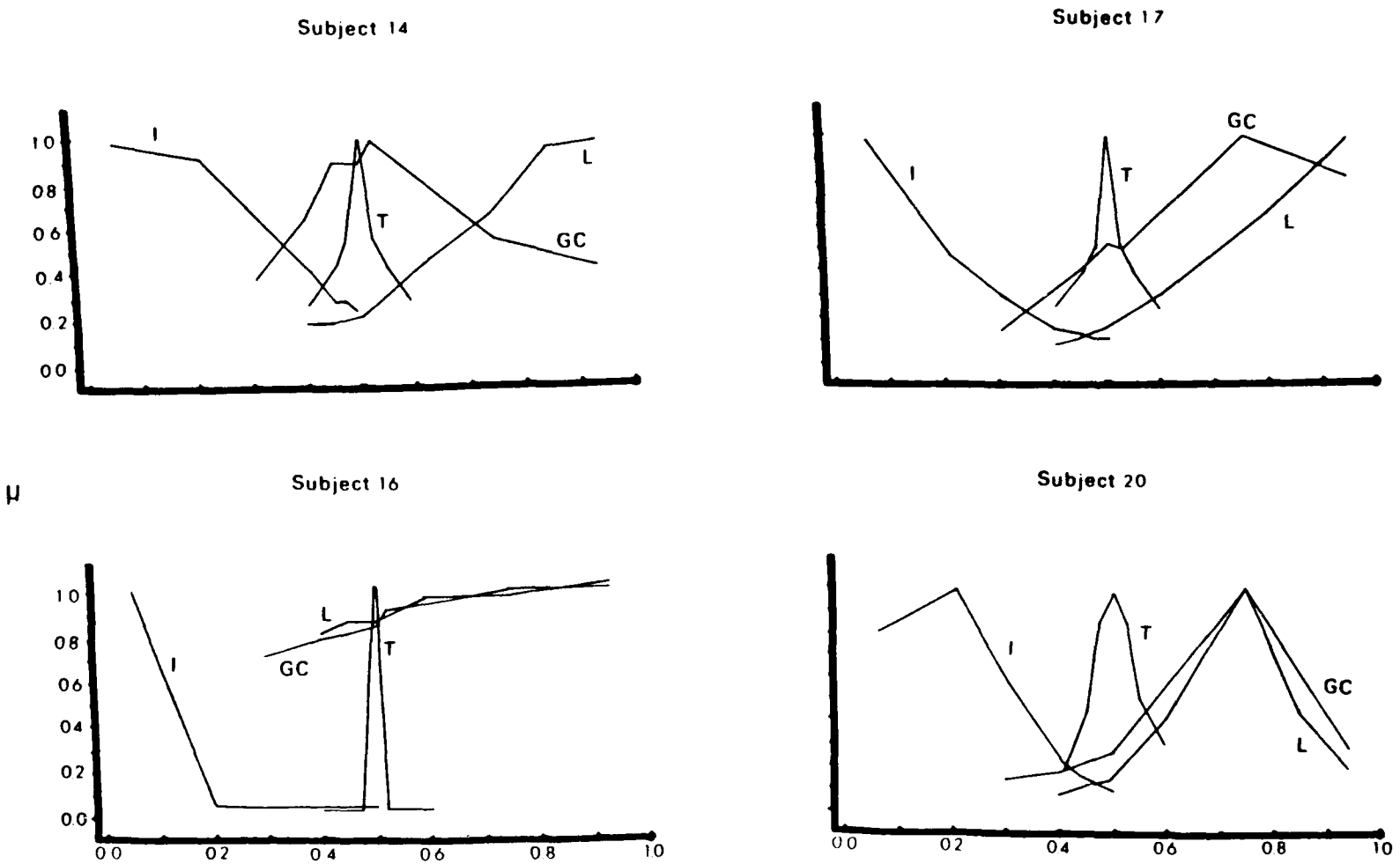
Subject 14
Subject 17
Subject 16
Subject 20
I
T
L
GC
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
μ
Probability

Fig. 10

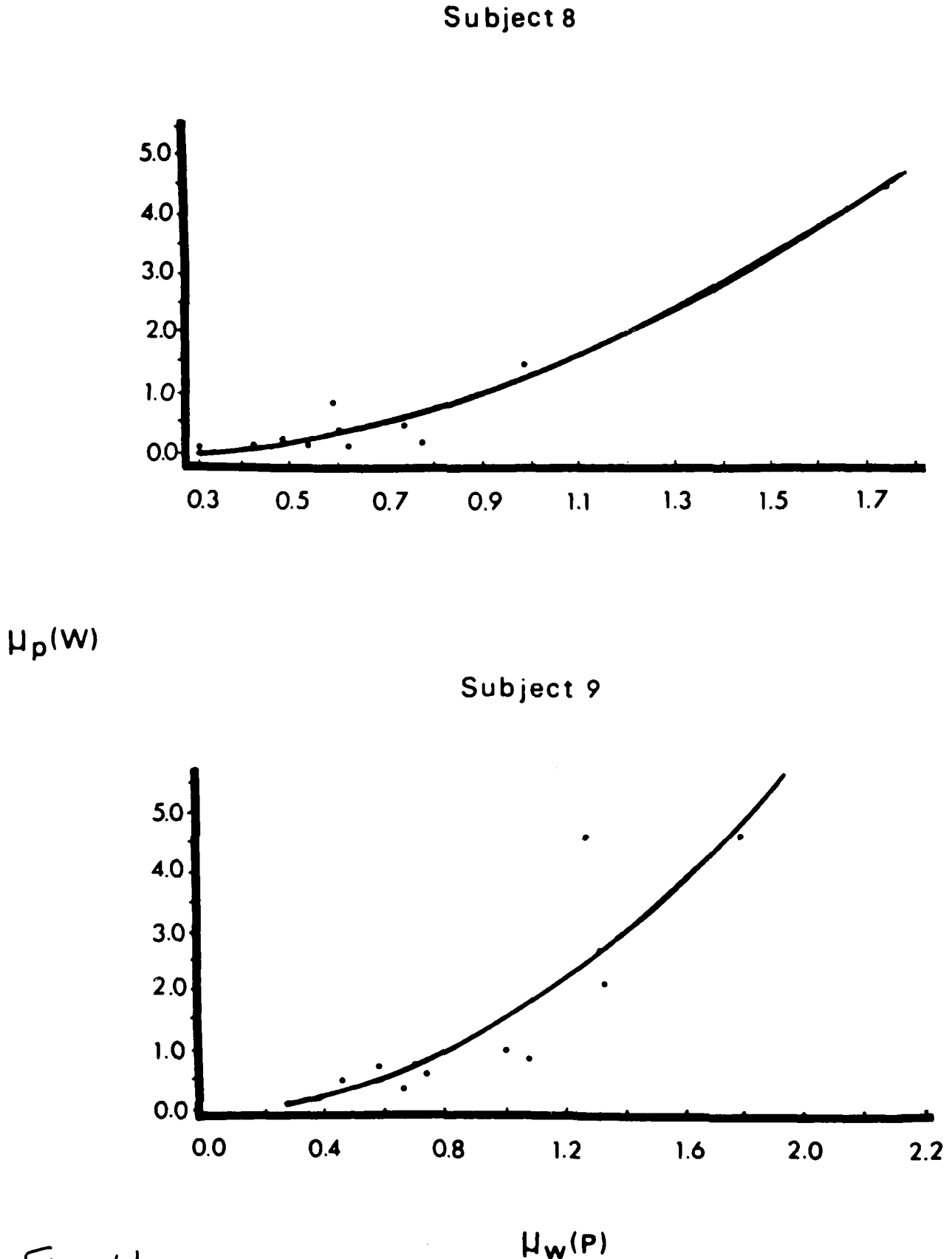
Subject 8
5.0
4.0
3.0
2.0
1.0
0.0
0.3
0.5
0.7
0.9
1.1
1.3
1.5
1.7
$\mu_p(W)$
Subject 9
5.0
4.0
3.0
2.0
1.0
0.0
0.0
0.4
0.8
1.2
1.6
2.0
2.2
$\mu_w(P)$

Fig. 11